Danish Natural Language Processing in Automated Categorization

Jonas Ekedahl <f99je@efd.lth.se>

March 17, 2008

Abstract

In this work, we investigate possible benefits of natural language processing tools, as means to support automated text categorization. Our corpus consists of a small collection of categorized Danish web pages in the fields of art, architecture, and design. The natural language processing techniques we examine are stop word removal, removal of functional words, and lemmatization. The tools are based on a stop word list, a part-of-speech tagger and a dictionary. We evaluate effects on a string matching classifier and a support vector machine.

The classification accuracy increases when using the lemmata, either in addition to or replacing the original inflected words in the documents. Positive effects are seen on both precision and recall. In absence lemmatization, the removal of stop words increases classifier performance, although not as much. Results are valid both for support vector machine, and string matching categorization.

Acknowledgments

First of all, I would like to thank my main supervisor Pierre Nugues, for guidance at all levels—theoretical, practical, as well as administrative. I am also very glad that Pierre allowed this project to have a start and me to continue with it for as long as it took.

I want to thank my co-supervisors Anders Ardö and Birger Larsen for inspiration and level-headed analysis and help in their respective fields. A special thanks to Anders for writing the Combine focused crawler, which enabled me to carry out this project.

I would also like to thank Koraljka Golub for advising me to write to Birger to begin with, Lars Nilsson—the guardian angel of computer science diploma workers, and all others who have supported me along the way.

Contents

Chapter 1

Introduction

According to Nugues (2006), *Natural Language Processing* (NLP) is "the automation of language processing by a computer." This "breed of linguistics and computer science" has, among many other things, developed tools whereby text can be analyzed with various applications. These may be segmenters of text into sentences or words, concordance extraction tools, spell or grammar checkers, annotators of words with their respective parts of speech, anonymization tools, information queries system, summarizers, or translators, just to mention some.

Automated text categorization is here to be understood as a supervised method to classify documents based on their text content (Golub, 2006). It consists of a three step method that involves the manual (hence supervised) classification of text documents into a set of topical classes, the derivation of class characteristics from the manually classified set, and the classification of new text documents based on those characteristics.

Categorization in this definition is to be seen in opposition to two related fields, namely classification and clustering. Clustering would be the unsupervised organization of documents into classes whose topics are not predefined, and classification the organization of documents into classes based on intellectually created controlled vocabularies (see Golub, 2006). The task can take the shape of the k number of binary classification problem such as in Aizawa (2001) or a less straightforward classification process where the features of the topical hierarchy can be incorporated in the classification algorithm, as in Koller and Sahami (1997).

Classification methods are often based on the presence or absence of predefined terms in the textual part of a document. The usage of NLP techniques would allow for connecting various forms that may appear different with a united semantic meaning, or separating terms that at a first glance seem identical but have different underlying meanings. This could be due to

things such as inconsistent vocabulary usage or different grammatical constructions.

1.1 Purpose

Keeping an organized collection of well-maintained and up-to-date web links in a world of dynamic electronic resources requires regular weeding of dead links, incorporation of new resources, and reanalysis of existing links as their content may change over time. Manually adding a new document to a collection, determining relevant keywords, and assessing the accurate subject topic are labor-intensive tasks that require relevant subject experts that are well educated in the particular collection standards. Needless to say, it is a costly endeavor. Hence, tools to support the classification process would allow for maintaining and developing organized digital information collections further.

The main goal of this thesis is to investigate whether there are benefits of utilizing NLP tools in assisting automated text categorization of web documents. The NLP methods we analyze are lemmatization and removal of words of little semantic content. These processes are based on a stop word list, a part-of-speech tagger and a digital lexicon.

1.2 Background

This Master's project is part of a degree project in computer science at the Lund Institute of Technology (LTH). The project started upon an idea from Birger Larsen, associate professor at Royal School of Library and Information Science, to use available NLP methods such as part-of-speech (POS) tagging, to assist in the manual classification task in Danish.

Tools to assist categorization of new documents into existing topical databases are needed for the maintenance and upgrading of services based on manually classified information. The field is a current topic of investigation, facing the unpleasant obstacles of real problems.

The objectives of this project is the creation and investigation of term lists for classifying documents in Danish based on NLP methods, applied to the existing on-line available Arkade portal (www.arkade.dk). The Arkade analysts form precisely such an organization that pay attention to advances in the field as they feel they might benefit from the outcome.

In Chapter 2, we describe the document collection used in the further experiments. The methods used are reviewed in detail in Chapter 3. The

results achieved and the conclusions to be drawn from them are explained in Chapters 4 and 5. In Chapter 6, ideas are given on how to explore the topic further. Chapters 7 contain general comments. We list the complete Arkade class structure in Appendix A, a few sample documents in Appendix B, the stop word list in Appendix C, and the two sets of word class tags in Appendix D and E.

Chapter 2

The Collection

2.1 General Overview

The Arkade portal contains 3,027[1] links to web documents covering various aspects of architecture, art, and design. The documents are manually classified into a hierarchical tree of classes, consisting of six main classes, further subdivided into a grand total of 322 classes spread over at most five hierarchical levels. Among the branches, there exist classes without further subdivision on every hierarchical level except for the top one. The documents can be member of classes at all levels in the tree. Each document may be categorized into several classes.

The documents were harvested by the Combine focused crawling system (Ardö and Golub, 2007) at three different occasions in order to get an instance of as many documents as possible. Most documents were crawled in the beginning of November 2006. As occasionally servers go down, we ran the last crawl in May 2007. This added 10 previously unseen documents. The first collected instance of every document is the one used in our further experiments.

2.2 Document Properties

2.2.1 Language

The documents are written in several languages, predominately English or Danish. We have chosen to look at only those documents identified as being written in Danish. After an inadequate attempt to identify automatically

[1] at the 24:th of July 2007

Table 2.1: Documents grouped according to language.

Language	Number of Documents
English	734
Danish	438
German	56
Swedish	50
French	24
Norwegian	13
Dutch	7
Italian	7
Finnish	5
Spanish	4
Estonian	3
Greenlandic	1
Japanese	1
Mixed languages	130
Unidentified language	4
Unclassifiable content	51
Irrelevant content	41

the language of each document, a manual examination was performed. Documents that are identified to have Danish as the main language with the exception of keywords or quotes in other languages are considered to be valid Danish documents.

The total document collection comprises 1,594 unique documents, 438 of which are considered to be written in Danish. Additionally, 72 documents are partly written in Danish. Others are either written in another language, or a mix of languages, or do not have a classifiable content (see Table 2.1).

2.2.2 Document Requirements

Another manual examination was performed in order to ensure that the quality level of the documents was satisfactory. All documents with fewer than six unique Danish words (including names) regardless of where (in title, plain text, or relevant meta tags), are considered too short for classification and are excluded. If the main message of the text of a document concerns technical issues regarding software or the internet, they are also excluded. This can be due to various error messages, automatic redirections, software requirements, or similar irrelevant information in regards to art, architecture, and design.

2.2.3 Document Distribution

In the end, the collection comprises 332 documents. This leave nearly 200 classes with less than two documents each (see Figure 2.1).

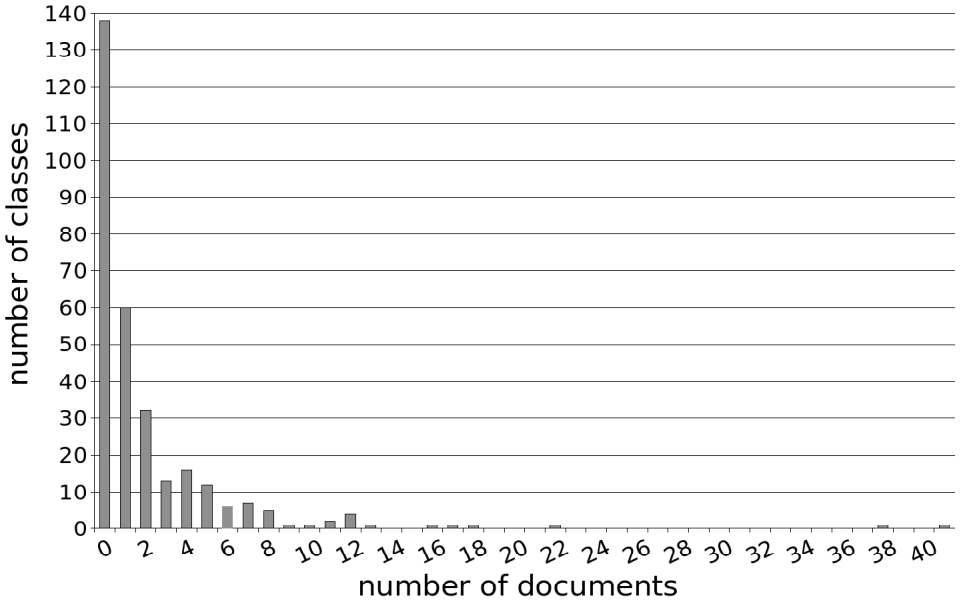

Figure 2.1: The spread of documents into the classes.

This distribution is far too sparse for an effective classification. The collection is obviously small in comparison to other well-known collections such as the Reuters Corpus Volume 1 with its ca. 8,000 documents per topical class on average (Lewis et al., 2004).

To overcome the major sparsity shortcomings, all documents were lifted to the top level of the hierarchy, leading to a total of six classes without hierarchical dependencies. As a consequence of merging classes, the average number of classes associated to a document fell to 1.28. Three quarters of the documents were members of only one class. Each document belonged to at most three classes, which was the case for only five documents. Figure 2.2 shows how many categories the documents were categorized into. Example of what documents look like after crawling are shown in Appendix B.

As Sun et al. (2003) point out, the relation between parent classes and child classes can either pose strong or weak subsumption constraints to the child classes in a hierarchical category tree. The lifting of documents is based on an unattested assumption that there is a strong subsumption constraint on a subclass from its parent class i.e., if a document belongs to a class it also belongs to the parent class(es) of that class.

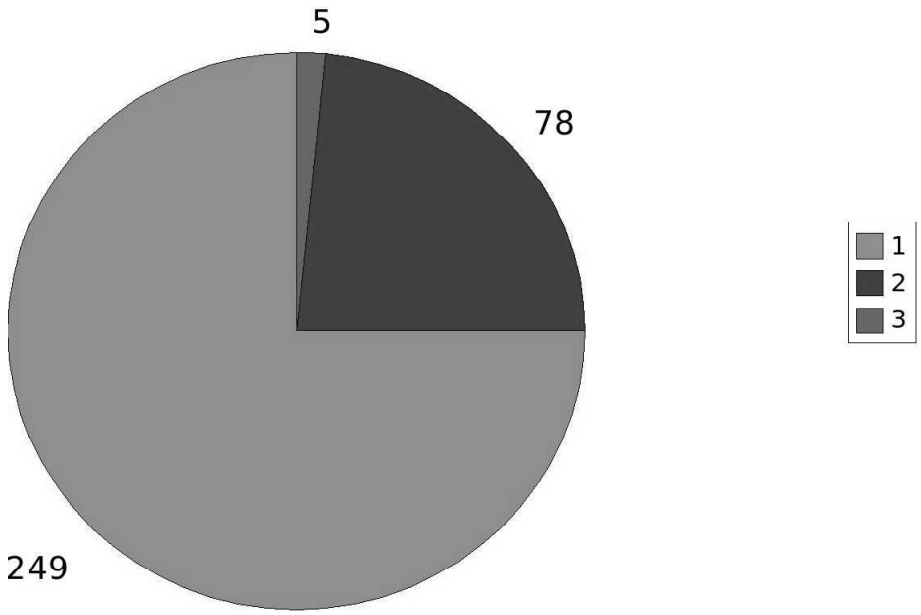

Figure 2.2: Documents sorted after the amount of classes to which they are associated

The distribution of the documents was hardly even. However, since all classes had more than 10 documents belonging to them, as can be seen in Table 2.2, it was deemed as sufficient.

At a later stage, when applying support vector machine categorization, the corpus was partitioned into mutually exclusive classes. This was conducted by assigning all double and triple classifications into their least common class, to ensure that all classes still were represented with at least 10 documents each. The size of the classes before reduction are shown in Table 2.2. This resulted in a drop in documents for the classes "Billedkunst" from 131 to 86, "Arkitektur, planlægning og byggeteknik" from 101 to 96, and "Kunsthåndværk og design" from 81 to 43 documents, as can be seen in Table 2.3. The others were not affected.

2.2.4 Terms

Ignoring capitalization, there is a total of 37,577 unique words in this collection. As seen in Figure 2.3 regarding the sum of the remaining documents, the distribution of words follows Zipf's law fairly well, as expected from a text collection, with a small number of very frequent words and an inversely proportional large number of rare words. The 10 most frequent words are listed in Table 2.4.

Table 2.2: The resulting six classes.

Class name	**Translation**	**Documents**
Billedkunst	Picture art	131
Arkitektur, planlægning og byggeteknik	Architecture, planning and building engineering	101
Kunsthåndværk og design	Crafts and design	81
Museologi	Museology	12
Konservering	Preservation	18
Kunstnere, fotografer, arkitekter, designere A-Å	Artists, photographers, architects and designers A-Z	77

Table 2.3: Reduced document distribution.

Class name	**Unique documents**	**Relative drop**
Billedkunst	86	34%
Arkitektur, planlægning og byggeteknik	96	5%
Kunsthåndværk og design	43	47%
Museologi	12	0%
Konservering	18	0%
Kunstnere, fotografer, arkitekter, designere A-Å	77	0%

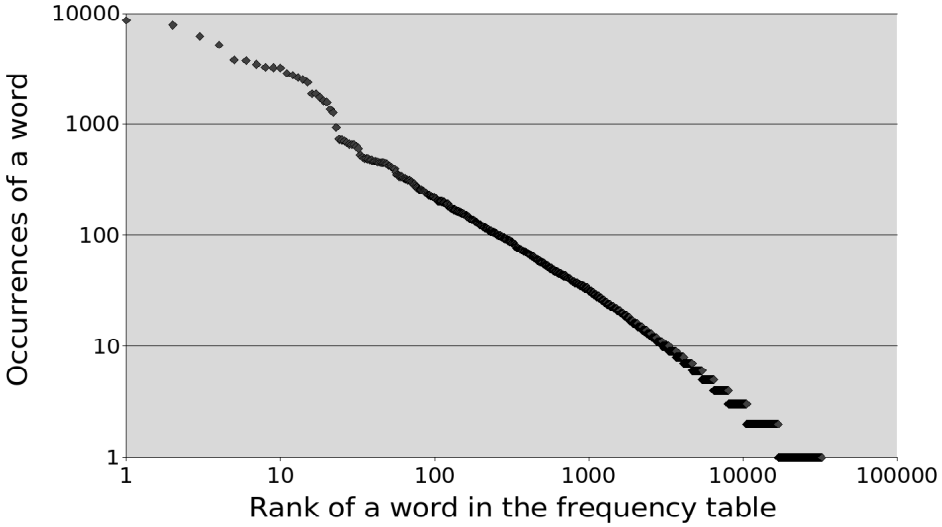

Figure 2.3: The word occurrences in the corpus ordered after falling occurrence—not too far from a straight line in a log-log diagram. Only the 32,500 most frequent words are shown.

Table 2.4: The ten most frequent words.

Word	**Translation**	**Rank**	**Total number of occurrences**	**Documents in which they occur**
og	and	1	8700	271
i/I[a]	in/you	2	7838	259
af	by/of/off	3	6253	235
er	am/are/is	4	5218	211
at	that/to	5	3793	164
en	a/an/one	6	3744	194
til	to	7	3464	224
for	because/for/went	8	3272	217
det	it/the/that	9	3240	172
på	on	10	3230	208

[a]As we will ignore capitalization we will not be able to separate the preposition *i* from the pronoun *I*.

Chapter 3

Method

In order to investigate the effects of different language processing tools on the performance of automated text categorization, a number of tools had to be implemented and connected. The experiment involved a tokenizer, a part-of-speech-tagger, a lemmatizer, classifiers, and measures for term representativeness and classification accuracy.

3.1 Language

As the bulk of the implementation is language dependent and as different languages need different treatment in their respective analysis, this project has focused exclusively on Danish. There are features in Danish that set it apart from many other languages. To our knowledge, texts in Danish have not been evaluated for automatic categorization. The Danish language has a rather high tendency for compounding and new compounds are created all the time. This leads to more topic-specific words. However, as their specificity increases, their generality drops and the low frequency of many words good for categorization could be problematic.

Danish is not a highly inflected language—however more so than English. Nouns are inflected in two cases, two numbers, and two degrees of definiteness. Verbs have different tense, voice, and mood forms, but neither gender, nor number. Adjectives are inflected according to three degrees of comparison, two degrees of definiteness, two numbers, two grammatical genders and different forms whether used adjectivally or predicatively. A lot of adverbs are constructed upon an adjective lemma to which they can be associated—in analogy with English *grateful* and *gratefully*. As a consequence, where lemmatization is useful, it might have an even greater impact on Danish than on English.

As a means to enhance the size of the collection, Norwegian documents could be included, as Danish and Bokmål Norwegian are rather close. This similarity has been utilized before such as in Bick and Nygaard (2007) and Henrichsen (2007). We did not implement it. It was deemed to have only a little benefit as there are but few Norwegian documents in our corpus. It would however introduce new kinds of errors.

3.2 The Tokenizer

The tokenizer is fairly straightforwardly implemented and is built on intuitive (naïve) assumptions about what words and sentences are, and unfortunately may split not only sentences but even concepts (such as *Dr. Rahardja*) at times. Documents may be structured to utilize pictures, animations, etc. as main information carriers. The written information may come in the shape of tables, concept listings, or scattered words rather than sentences, depending on document design. These are ignored by our setup.

- The text is split into sentences where every dot, question mark, exclamation mark or carriage return indicate the beginning of a new sentence. Those symbols are not considered parts of words.

- The symbols _ , & , \ , /, ', +, # , ~, %, =, ^ , *, ‘, <, >, |, @, $, ", nine other nonword characters, and tabs are considered invalid parts of words and are treated as blank spaces.

- Commas, colons, semicolons, the double quotation marks ", and the brackets (,), {, }, [, and] are kept as sentence segment separators, but are not considered as parts of a words.

- Series of blanks (including other signs treated as blanks) are treated as single blanks.

- Numbers are kept in order not to confuse the part-of-speech tagger, but are not considered valid words, despite the semantic content of particular years etc.

The tokenizer assumes that the text it gets as input is relevant, which is not the case everywhere in the collection. Some erroneously written HTML code will contaminate the text with HTML tags, server messages and/or other undesired features, which may stay embedded in the tokenizer input.

3.3 The Part-of-Speech Tagger

In order to identify the function of words in their context, such as separating nouns from verbs, we used the Stanford part-of-speech tagger (Toutanova and Manning, 2000; Toutanova et al., 2003). It was trained on the part-of-speech annotated part of the Danish PAROLE corpus (Keson, 2007). This consists of sentences from a mix of sources, 250,209 part-of-speech annotated words in all.

It was observed that the part-of-speech-tagger did not perform equally well, on all material. As expected, it performed better on text segments of full sentences and less well on text segments consisting of scattered words. No thorough examination was conducted to measure its accuracy. A brief examination however gave an estimate of about 94% correct word class prediction on words. The test was done on flowing text, preprocessed by our tokenizer. Our rough estimations are acceptable and not too far below the 97.24% per-position tag accuracy that Toutanova et al. (2003) achieved. As further tools do not use the information of the particular inflectional form produced by the part-of-speech tagger, no examination was conducted to examine the accuracy beyond the basic word class tags.

3.4 The Lemmatizer

A dictionary was constructed upon a list of Danish words and their respective inflections (Corpus, 2007). It consists of ca. 80,000 stems with information on word class and the complete set of inflections.

The dictionary takes a lexeme and its corresponding word class. From them, an uninflected form of that word is returned—a lemma (or lemmata in the case of multiple possible interpretations).

There are a few problems with this lemmatization:

- Only a fraction of all Danish words are present in the dictionary and subsequently there must be a rule to handle out-of-dictionary events. In this case, our choice landed upon simply returning the lexeme itself. An alternative could be a rule-based approach, which is a tool in itself. For a wider overview, consult Jacquemin and Tzoukermann (1997).

- Although the word exists in the dictionary, the parser tag might not match any of the dictionary entries, due to erroneous parsing or an out-of-dictionary word usage. The latter is probably the less likely. A set of rules on how to handle all possible combinations of parsed tag vs. tags in the dictionary is required. Analysis of input and output for a few

Table 3.1: Inflections of the nouns *læg*.

		indefinite		definite	
stem	**number**	**nominative**	**genitive**	**nominative**	**genitive**
$læg_1$	**singular**	læg	lægs	lægget	læggets
	plural	læg	lægs	læggene	læggenes
$læg_2$	**singular**	læg	lægs	læggen	læggens
	plural	lægge	lægges	læggene	læggenes

different samples posed a strong argument towards simply neglecting the part-of-speech tag and incorporating all lemmata obtainable from the dictionary that are of a relevant word class. This is the strategy we selected and implemented.

- Several lemmata may be identical in form. Compare the inflection of the nouns $læg_1$ 'sewn fold' and $læg_2$ 'calf' in Table 3.1.

  Although a word form such as *læggets* only may exist as an inflected form of $læg_1$, no action has been undertaken to keep lemmata identified or separated, as it has been deemed having only a minor effect, as well as go beyond the capabilities of the chosen classifiers.

- Removal of words from word classes not interesting for classification purposes, such as pronouns or prepositions, should be conducted carefully. Two factors could cause problems:

  1. If there is any risk of such functional words slipping through the lemmatizer, for instance wrongly tagged as adjectives, they are suddenly a lot fewer in the corpus than before. In a difference to very frequent words, scarce words are hard to suppress through distribution analysis measures in later stages. If one cannot make sure that all instances of a word are removed, it might be better to let them all slip through.
  2. Removing functional words will also affect other potentially relevant terms with the same identical word form. Attempt to further remove the lemmata responsible for the clashes might likely lead to more clashes between further word forms of new lemmata. For instance, if all words spelled[1] *blot* are removed because of the

[1]The pronunciation might differ depending on the meaning.

subordinating conjunction[2] *blot* 'if only', the distribution of the adjective *blot* 'mere', adverb *blot* 'only', the noun *blot* 'sacrifice', and the verbs *blote* 'sacrifice' and *blotte* 'uncover' would all be affected. They would all be less frequent, but still present[3], and in an altered distribution. Removing all of them all together would in turn affect the noun *blotter* 'flasher', as *blotter* is the present tense form of *blotte*, and the adjective *blottet* 'exposed', as *blottet* is the supine of *blotte* etc.

By necessity, there is a trade off between the two factors above. In our setup we have chosen to trust the Part-of-speech tagger and ignore factor 1 above.

3.5 Stop Word Removal

Upon making a simple tool to remove the most frequent words the choice fell on the Danish stop word list used by Porter (2002). It consists of 94 words, most of them being pronouns, prepositions and/or conjunctions. The ten most common ones are shown in Table 3.2. The entire stop words list is shown in Appendix C.

3.6 The Selection of Terms

3.6.1 Relevance Measure

We choose to use a $tf \times idf$ measure (Salton and Buckley, 1988), to extract the relevant terms for the description of the classes, from the document collection. The term $tf \times idf$ reads Term Frequency times Inverted Document Frequency. This measure is to represent the commonness of a feature in a subcollection in regards to the commonness of that feature in the considered document universe. The $tf \times idf$ measure comes in various shapes and forms. There are also other measures that could have been used, such as the *odds ratio* (Brank et al., 2002), or *Rep(·)* (Hisamitsu et al., 1999). The $tf \times idf$ is known to work well in text categorization and "is one of the most commonly used term weighting schemes in today's information retrieval systems" (Aizawa, 2003). Aware of what (Robertson, 1990) formulates as:

[2]These words are somtimes referred to as complementizers.
[3]The adverb *blot* is an exception since it has no other form.

Table 3.2: The ten most frequent stop words.

Stop word	Translation	Occurrences
og	and	8700
i	in	7838
af	of/by/from/off/for/in/with/on, off	6253
er	present tense of "to be"	5218
at	that (in front of a sentence)/to (with infinitive)	3793
en	a/an	3744
til	to/at/for/until/against/by/of/into, more	3464
for	at/for/to/from/by/of/ago, in front/before, because	3272
det	that (dem. pronoun)/it (pers. pronoun)	3240
på	on/upon/in/on/at/to/after/of/with/for, on	3230

> "a term weighting formula that provides appropriate weights for use in a match function for retrieval is not necessarily an appropriate measure for term selection in the first place",

term selection has nevertheless been introduced through the cropping of low $tf \times idf$ terms from out of the term lists.

In accordance to Baeza-Yates and Ribeiro-Neto (1999) the $tf \times idf$ measure we utilize is defined by:

$$weight_{t,d} = \left(0.5 + \frac{0.5 \cdot t_{i,d}}{max_l(t_{l,d})}\right) \times \log\frac{n}{x_t} \tag{3.1}$$

where:

- $t_{i,d}$ denotes the frequency of term i in document d
- n denotes the number of documents in the collection
- x_t denotes the number of documents where the term t occurs.

3.6.2 Inclusion Policy

The relevant document universe is in this case the set of documents that are, or ideally should be, incorporated into the Arkade portal now or at some point

in the future. The latter is hard to measure directly, so currently associated documents are taken as a representative sample, from which features can be extracted under the assumption that they significantly well represent the document universe as a whole.

We have chosen to include all words regardless of how frequent they may be in the corpus. Although it is statistically hard to motivate the association of a sparsely distributed term to one or more classes, according to Aizawa (2001), the information gain, through the collection of sparse terms as a whole contributes significantly to the information available to separate between the classes—"use as many terms as possible." The inflation of the feature space due to an inclusive approach towards terms is not a problem that cannot be overcome in this relatively sparse setup.

One could have incorporated some terms with negative weights, whereby their presence would make the class associated to such a term *less* likely. Due to the small size of the collection in relation to the number of classes, negative weights have not been utilized, as this feature is even more dependent on the extensiveness of the document collection, and thus deemed not applicable in this setup.

3.6.3 Shortcomings

Quite a few documents contain some words in a language other than the main language of the document. This may be due to such things as foreign quotes or keywords in the text, standardized page layout regardless of chosen language, or bad HTML programming resulting in HTML tags being embedded in the text. This leads to that a few alien words may be considered very specific for their class despite the fact that they may be functional words or other words not well suited for class description or differentiation.

The lemmatization process may also worsen the error by interpreting the foreign words as inflections of Danish words and thus expand the class features with misplaced correct Danish lemmata. For instance the (English) HTML-tag *head* could be interpreted as an imperative of the Danish verb *heade* 'head' without guiding context.

It is normally advised to focus on complex terms such as multiwords and compounds. Single words are often ambiguous. Categorization based on complex terms normally performs better than classification based on the same number of simple terms (Aizawa, 2001). As the complexity of terms grows so does the demand for an exhaustive corpus. In the corpus there are a lot of possible complex terms, most of them infrequent. Expaning the feature space with complex terms would inflate the feature space largely. As a great deal of the complex terms are unique, the feature space enlargement

would be largely in vain. As one-word-terms are very straightforward to use, only they were used initially. The call for more complex terms was at later stages not deemed crucial and the feature was not further explored.

3.7 The Classifiers

The semantic content of a document is carried by the content of the individual words of the document, their interrelated meanings and the association they induce in the reader. The meaning of the document can thus be modeled by the collection of separate individual words that it contains. Therefore words and/or lemmata are suitable features when analyzing the content of a text document. This model is called the bag-of-words approach.

3.7.1 The String Matching Algorithm

Description

The integrated automated topic classifier in Combine (Ardö and Golub, 2007) is a string matching classifier. It can achieve results that are comparable to those of state-of-the-art machine-learning algorithms for certain classes and applications (Golub, 2007). It is built on the bag-of-words approach, and categorizes documents by comparing terms extracted from the document with the terms from a predefined term list.

Extract from term list: *(weight: term=class code)*

33: forvaltningarkiv=01/010
5: dokumentation=01/010
5: kunstdatabasen=01/010
5: arkitekturforskning=20/210/2105
1: designnetværk=50/510
7: overfladebehandling=70/730
5: terpentinolie=70/730

To each word in the word list there is a class and a weight associated. For every match between the document and an instance of a term in the term list, the documents association to that class increases by the weight. All weights associated to a specific class are summed up. The result is a list of classes with assigned weights. The weights are to represent the relevance of the terms are in the document:

- The more of the relevant terms there are, the higher the weight.
- The higher the term list weights are for the terms present in the document, the higher the document/class pair is weighted.
- The terms with a more significant position i.e., closer to the beginning of the text, are given a higher score.

The score S of a class is calculated as follows:

$$S = \sum_i \sum_j tw_i \cdot lw_j \cdot f_{ji},$$

where tw_i is the weight associated to term i, lw_j is the weight associated to the location j in the document, and f_{ij} is the number of occurrences of term i at location j.

The result for a document is a list of classes with corresponding weight sums. An example what a class list might look like is given below in the format *code="class code" score="weight sum"*:

code="70/740/7401" score="819"
code="20/260/2603" score="234"

The current document is likely to cover rather the topic "Photographic documentation" connected to the code "70/740/7401", than "Cities and districts" connected to the code "20/260/2603".

Upon excluding the classes that score low, we produce a categorization.

Preprocessing of the Text

As seen above, Combine goes beyond the basics of the bag-of-words approach. In the chosen implementation, it considers not just the number of term occurrences, but also their position in the text—the earlier a term occurs, the more significant it is considered to be. As the textual information originates from different sources within a single document, it is thus important to weigh which of these are more significant and thereby should be sent first to the classifier.

Apart from the bulk *plain text* segment, the meta tags that are taken into consideration here are the ones the Combine crawler delivers as *description*, *subject*, or *title*, in the Dublin Core metatag standard (Weibel et al., 1998; Ardö and Golub, 2008). The *title* is only considered when it differs from the HTML title of the document, which is included in the *plain text* segment.

These are then concatenated in the order *title*, then *description* and *subject*, and finally *plain text*, thereby having a unified text segment where the parts are arranged so that the more a word or a sequence of words are likely to be relevant for the classification process, the sooner they will appear in the segment, as is empirically supported by Golub and Ardö (2005).

3.7.2 Support Vector Machines

If every word or lemma would constitute one semantic dimension, there would be a feature space in which all documents can be placed—the span the total set of independent dimensions, each consisting of a word/lemma in at least one document in the collection.

What a support vector machine (SVM) (see Boser et al., 1992; Cortes and Vapnik, 1995) does is that it assigns separating hyperplanes to a multidimensional feature space. These hyperplanes work as discriminators between instances that share and that do not share a particular feature (see Figure 3.1). In other words, a hyperplane creates a grouping function for an unknown but desired feature, hence it is a classifier into two disparate groups.

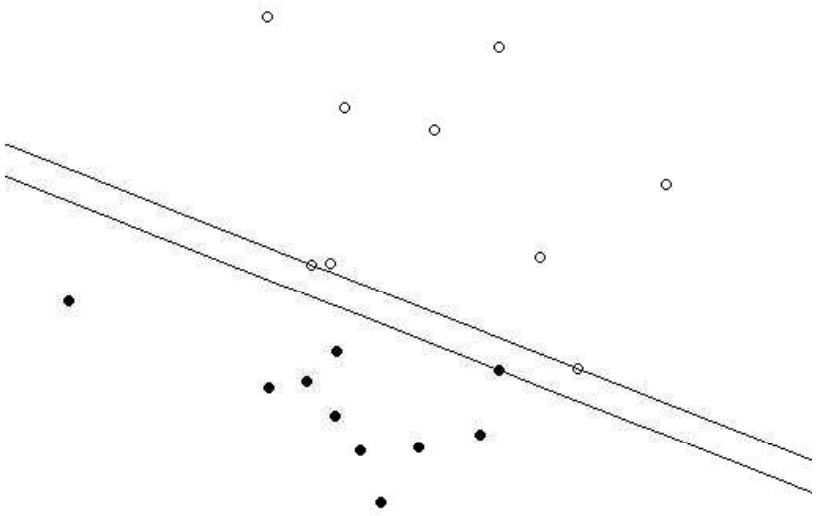

Figure 3.1: A two-dimensional illustration of a maximized margin discrimination between two sets of elements—filled versus open circles.

If several hyperplanes are calculated, discrimination into more than two

categories can be achieved. Also, more hyperplanes allow for separation between more complexly distributed training instances, in the feature space.

It is common to reduce the feature set by associating different terms with related semantic content to the same feature dimension. This is done in order to associate text segments with different writing styles and/or vocabulary use to the same semantic content, remove words whose meanings are irrelevant to the particular topics involved, and speed up the computation, which usually is highly dependent on the number of features involved. This is a technique that was utilized in these experiments in the form of the association of a lemma to several inflected words forms. This is done where it is explicitly stated that categorization was performed with lemmata only.

The way we convert a document into its vector representation, every instance of a word adds one to the feature dimension that represents it. After that, the vectors are normalized in such a way that every dimension gets a value between 0 and 1. This is done simply by dividing the values of all dimensions with the highest dimensional value in that particular document.

Sequential Minimal Optimization (SMO) is a SVM utilizing a minimalistic approach towards the calculation of hyperplanes, allowing for fast calculation that can compensate the minimalistic approach by sheer computation volume (Platt, 1998). The SVM we use is the standard SMO in the Weka software package (Witten and Frank, 2005). SVMs are shown to work well in the field of text categorization, as is stated by Joachims (1998). More information on SVMs, the theory and mathematics behind the SVM, the SMO, and text categorization with the help of SVMs can be found in Cristianini and Shawe-Taylor (2000).

3.8 The Categorization

Categorization is an intrinsically difficult task. In terms of manual categorization, "a review if the literature reveals a rather weak level of interindexer consistency whatever the context" (David and Giroux, 1995).

Categorization is a task that according to David and Giroux (1995)

> "... can be characterized as an ill-defined problem inasmuch as, despite certain established indexing procedures, the nature of the goal is defined by the indexer".

We will not examine to what degree the associated classes to documents are valid and/or informative for a user. We rather assume that the classification done at Arkade.dk is the gold standard, whose properties we shall seek to resemble.

We sent the document collection through a pipeline NLP processes, whereby the content was transformed into a set of terms. These terms, representing the documents, were then used to train and evaluate a classifier.

3.8.1 The NLP Pipeline

Pipelines of NLP tools were put together. These may include the following processes: tokenization, stop word removal, part-of-speech tagging, lemmatization, and removal of functional words. Lemmatization is here to be understood as either the adding of lemmata or replacing with lemmata—both were used. The documents, fulfilling the requirements described in Section 2.2.2, were sent through the chosen pipeline, and the outcome was sent to a classifier.

The categorization by the classifier was conducted as a five-fold cross-validation. It means that the documents are split into five equally large sets. A training collection is formed comprising four of these and hence about 80% of the documents. The last piece forms a test set. Assembling four out of five pieces can be done in five different ways and accordingly, we receive five set pairs, each containing a training set and a test set.

The chosen classifier is then trained on a training set and is then set to categorize the corresponding test set. The training and testing of the classifier is repeated five times–one for every set pair. Up to ten different NLP pipelines are evaluated and hence there are up to 50 different training and testing instances.

1. Tokenization
2. Tokenization, Stop word removal
3. Tokenization, Part-of-speech tagging, Addition of lemmata
4. Tokenization, Part-of-speech tagging, Addition of lemmata, Stop word removal
5. Tokenization, Part-of-speech tagging, Addition of lemmata, Removal of functional words
6. Tokenization, Part-of-speech tagging, Addition of lemmata, Removal of functional words, Stop word removal
7. Tokenization, Part-of-speech tagging, Replacing with lemmata
8. Tokenization, Part-of-speech tagging, Replacing with lemmata, Stop word removal

9. Tokenization, Part-of-speech tagging, Replacing with lemmata, Removal of functional words

10. Tokenization, Part-of-speech tagging, Replacing with lemmata, Removal of functional words, Stop word removal

All pipelines include the tokenization. As the lemmatization and the removal of functional words require input text to be part-of-speech tagged, where we mention that lemmatization or removal of functional words has been conducted it is implied that part-of-speech tagging has already been done.

3.8.2 Representation of the Documents

Based on a training set, the frequencies of individual words in documents and the topical classes of documents are calculated. Word may refer to word and/or lemma. For rapidity, each document is represented as a column in an inverted index. As a whole, this index thus includes all words present in the training set and the values describing their frequencies in each document and in every topical class.

From this index, a term list containing all words deemed to be significant enough to associate to a class is assembled. Each word is put together with a value stipulating just how strong the word associates to its particular class. As only positive weights were being used, only words whose distribution is significantly higher in one specific class were listed, in a difference to significantly lower frequencies of terms for one class that could be included and assigned negative weights.

By comparing the achieved categorization with the placement of the documents in the original Arkade tree structure, we can measure the described categorization procedure. As we initially have different preprocessing sequences, we get categorization results for different setups using different NLP approaches. Thereby we can measure the impact of NLP tools on automated text categorization.

3.9 Evaluation and Error Analysis

While comparing the different classification results, it is of essence to use evaluation measures that well capture the intent and purpose of the research. The following metrics were implemented: precision, recall, and F_1 measure. All measures may then be micro- or macro-averaged on both class and document level.

From the two sets A and B, where B is a prediction of A, the *precision* is the ratio of elements in B found in the corresponding place in A and the *recall* is the ratio of elements in A that are correctly predicted in B. If α is to denote all documents correctly categorized into their respective classes, β is to denote all documents categorized into wrong classes, and γ is to denote all documents not categorized into their right classes, formulae are as follows:

$$Precision = \frac{\alpha}{\alpha + \beta}$$

$$Recall = \frac{\alpha}{\alpha + \gamma}$$

In our case, we have a set of sets of predictions, namely the set of classes each into which a set of documents are associated. Alternatively we have the set of documents, and each document has a set of classes associated. Our measures can thus be normalized in different ways:

- If we just consider the total set of predictions in which a document has been associated to a class, the ratio of such correct predictions out of all predictions is the micro-averaged precision. The micro-averaged recall is analog.

- The classwise macro-averaged precision (CMP) is here defined as the average of the precision of every class and the classwise macro-averaged recall (CMR) analogously:

$$CMP = \frac{\sum_{i=1}^{n} \frac{\alpha_i}{\alpha_i \cdot \beta_i}}{n}$$

$$CMR = \frac{\sum_{i=1}^{n} \frac{\alpha_i}{\alpha_i \cdot \gamma_i}}{n},$$

  where i denotes the class, and n the total number of classes.

- The precision for a single document can thus be calculated as the ratio of classes associated to the document that is correct. Averaging the results over all documents leads to the document-wise macro-averaged precision (DMP). The document-wise macro-averaged recall (DMR) is defined in an analog manner.

$$DMP = \frac{\sum_{i=1}^{n} \frac{\alpha_i}{\alpha_i \cdot \beta_i}}{n}$$

$$DMR = \frac{\sum_{i=1}^{n} \frac{\alpha_i}{\alpha_i \cdot \gamma_i}}{n},$$

where i denotes the document, and n the total number of documents.

The F_1 measure is the harmonic mean of the corresponding precision and recall:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}.$$

There are thus a classwise macro-averaged, a document-wise macro-averaged and a micro-averaged F_1.

In the case of five-fold cross-validation (see Section 3.8.1) the categorizations of all splits are summed up before evaluation metrics are calculated, as if all of documents had been categorized at once.

The classwise macro-averaged precision in the case of k-fold cross-validation is hence calculated as:

$$CMP_k = \frac{\sum_{i=1}^{n} \frac{\sum_{j=1}^{k} \alpha_{i,j}}{\sum_{j=1}^{k} \alpha_{i,j} \cdot \sum_{j=1}^{k} \beta_{i,j}}}{n},$$

where n is the number of classes and k the number of splits. $\alpha_{i,j}$ is thus the number of documents that have been correctly categorized into class i in split j and $\beta_{i,j}$ the number of documents that incorrectly have been categorized into class i in split j.

Chapter 4

Results

4.1 String Matching Categorization

4.1.1 Full Class Structure

Two different setups are made. In the first one, we categorized tokenized documents. This is conducted upon a term list of unprocessed *raw* words, found in a test set—*Word forms*. In the second one, documents as well as term lists were complemented with the lemmatized word forms where those could be found, here named *Lemmata and word forms*.

The document set were split into two sets—a training set and a test set—of the relative sizes 90% vs. 10%, so that the bulk of the documents were used to supply the classifier term lists with relevant terms.

Out of all evaluation measures but one, the addition of lemmata outperforms classification without lemmata (see Table 4.1).

It is clear that the current setup is not sufficient in producing relevant advice even as an assistance in a manual classification task, let alone as a

Table 4.1: Full class string matching results.

	Word forms	**Lemmata and word forms**
Assigning precision	2.6%	**2.8%**
Assigning recall	33.3%	**35.7%**
Class precision	2.5%	**3.0%**
Class recall	32.6%	**35.9%**
Document precision	**3.1%**	2.9%
Document recall	37.8%	**39.9%**

Table 4.2: String matching categorization with full word forms.

Word forms	**Split 1**	**Split 2**	**Split 3**	**Split 4**	**Average**
Assigning precision	2.6%	3.1%	2.5%	2.9%	2.8%
Assigning recall	33.3%	45.7%	31.0%	48.6%	39.7%
Class precision	2.5%	2.5%	2.2%	3.3%	2.6%
Class recall	32.6%	42.9%	26.7%	45.6%	36.9%
Document precision	3.1%	2.0%	2.4%	2.9%	2.6%
Document recall	37.8%	45.5%	34.5%	47.8%	41.3%

reliable classifier for usage in other applications. Largely this is due to insufficient material from which to draw conclusions. Incorporation of additional manually classified Danish documents is not possible however.

As the combined term list of both lemmata and word forms is rather heavy computationally few total investigations were conducted.

To verify whether these results are typical several different 90/10 splits and corresponding classifications of the *Word forms* kind was performed, as can be seen in Table 4.2.

The *Split 1* that corresponds rather well to the average forms basis for the examination of also the addition of lemma forms presented above.

4.1.2 Reduced Class Structure

The sparsity of documents in regards to the rather large set of categories (322) leads to rather poor performance results. To overcome this problem we construct a setup where only the six top classes are considered (see Section 2.2.3). If not proper thresholds are applied, documents could easily be categorized into all classes, as there are only six classes available. Therefore we put up a limit on the number of categories into which a document is allowed to be categorized. For simplicity, we set it to one—the one deemed most accurate by the classifier. For a perfect categorization three classes per document are required, as there are documents that belong to up to three classes. This means that perfect categorization is an impossibility. Otherwise, the string matching categorization in the new setup is perfectly analog to the one conducted above.

Analysis for Fix Term List Size

The number of terms associated to each class was set fix to 30. These terms are chosen based on $tf \times idf$ values. The weight associated to them in the

Table 4.3: String matching categorization over six classes.

Term type(s)	**Filtering out**	**Micro-averaged F1**	**Macro-averaged F1**
Word forms	nothing	10.2%	14.9%
Word forms	stop words	10.4%	**17.8%**
Lemmata and word forms	nothing	14.1%	**20.4%**
Lemmata and word forms	stop words	13.9%	19.1%
Lemmata and word forms	functional words	13.0%	18.4%
Lemmata and word forms	functional and stop words	13.1%	15.9%
Lemmata	nothing	14.3%	**21.1%**
Lemmata	stop words	14.0%	19.5%
Lemmata	functional words	15.0%	18.8%
Lemmata	functional and stop words	14.5%	18.8%

string matching categorization was the $tf \times idf$ rounded down to nearest integer value larger than zero.

We do not experiment with removing functional words in combination with not performing lemmatization. The filter for functional words requires the heavy part-of-speech tagging procedure. It does not seem meaningful to do the part-of-speech tagging and not use the results for identifying the lemmata of the words, either as complement or replacement of word forms.

As can be seen from Table 4.3 the results are overall rather poor. They are a lot better than when using 322 categories but far from usable or satisfying. It seems beneficial to use lemmata for categorization either alone or as a complement to the word forms in the documents. The removal of functional words seems to have a negative effect. Stop word removal surprisingly seems to have a slight negative effect with the exception of the case of no lemma extraction.

The Influence of Number of Terms Used

It is often sought for precision and recall to be near equal. The $F1$ metric often tends to reach a maximum as precision and recall intersects. As results so far show rather different value to precision and recall it is of special interest to investigate additional parameters in order to achieve better categorization. The most obvious is to vary the amount of terms per class. The initial choice

of 30 was an ad hoc assumption.

Three different setups were further analyzed following a doubling in used terms in the categorization term lists until the training stage for at least one class could not provide sufficient term numbers. This size varied with preprocessing type as well as with the different training set splits.

Analyzing the results show that results indeed depend on number of terms and that $F1$-measure for 30 terms per class is a lot lower than the maximum (see Figure 4.1). There seems to be a threshold beyond which performance significantly drops.

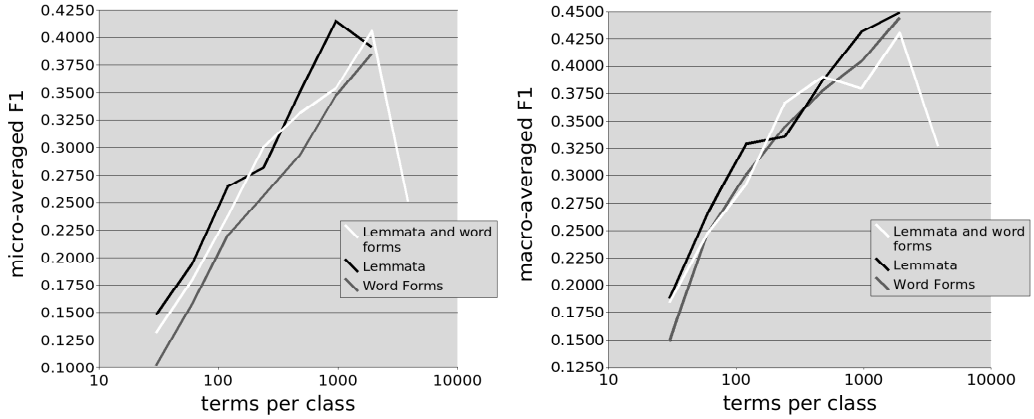

Figure 4.1: Micro-averaged F1 value as a function of terms per class during string matching categorization.

The increased performance is due to increased recall as more terms are considered. The increased number of terms seems to have only modest impact on precision up to a certain point after which both precision and recall falter (see Figure 4.2). The threshold is characterized by a sudden categorization of most documents into only one class. Why this comes about is not fully examined.

The downward slope occurring at the last measuring point (3840 terms per class) has a direct correspondence also in the unprocessed categorization in the cases when it was possible to achieve such an amount of terms per class. For the maximally reduced categorization, it was not possible to achieve 3840 terms per class in any split. It was thus not verified whether the same thing happens in this case.

As a comparison, "categorization" by simply assuming that all documents would belong to the most frequent class would yield a macro-averaged F1-value of 23.9% and a micro-averaged $F1$-value of 34.7%. To outperform both this micro-averaged F1 baseline, this classifier needs to be optimized in

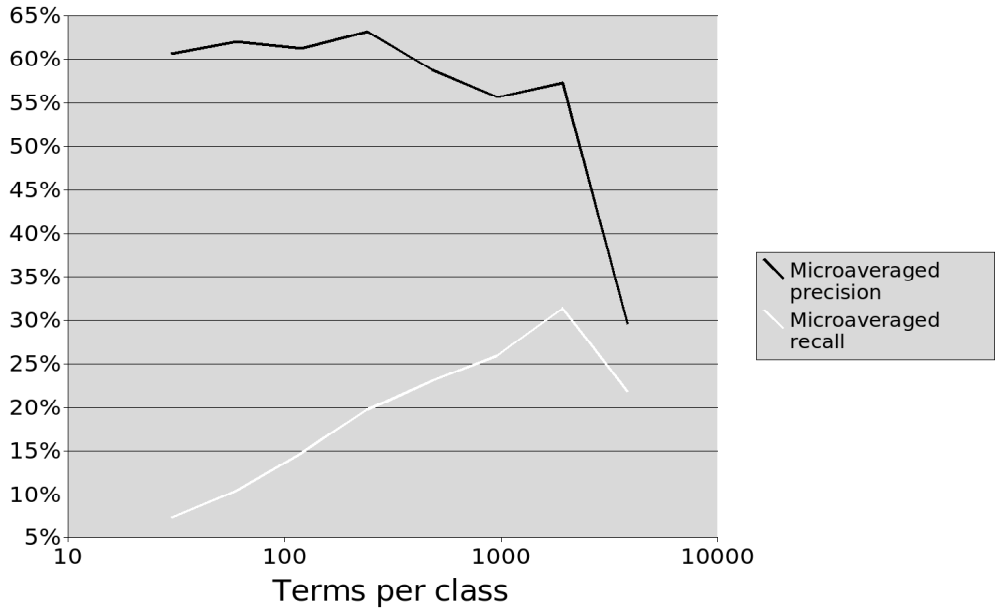

Figure 4.2: Micro-averaged precision and recall as a function of terms per class during string matching categorization with both word forms and their corresponding lemmata, from which functional words but have been removed, but no stop word removal. The tendency of how precision and recall vary depending on the number of terms seems to be general.

regards to terms per class.

4.2 Support Vector Machine Categorization

In order to achieve more general conclusions, we conducted a second experiment based on support vector machines. The documents were preprocessed by sending them through a pipeline of processing tools. The sets contained either the words in the form present in the document whatever the inflection, the lemmata of the words in the document or the combination of the particular inflected form and its lemma. In addition, in the cases where lemmatization had been conducted the terms could be removed, if identified as belonging to a class of words of little semantic content. In either alternative, the documents could either be sent through a stop word filter or not.

The different alternatives lead to 10 different document sets based on amount and type of preprocessing. Each set was then transferred into an inverted index of the Weka arff format (see Witten and Frank, 2005). To sim-

Table 4.4: Micro-averaged support vector machine categorization.

Term type(s)	**Removal of functional words**	**Stop word removal**	**Micro-averaged** $F1$
Word forms	No	No	53.6%
Word forms	No	Yes	56.0%
Lemmata and word forms	No	No	56.0%
Lemmata and word forms	No	Yes	**57.2%**
Lemmata and word forms	Yes	No	56.3%
Lemmata and word forms	Yes	Yes	**57.2%**
Lemmata	No	No	56.9%
Lemmata	No	Yes	56.6%
Lemmata	Yes	No	56.0%
Lemmata	Yes	Yes	56.9%

plify the process each document was considered to belong to only one class. In the case of documents belonging to several classes, the classes that were most frequent in the corpus were removed, thereby ensuring a certain minimum of documents per class and achieving a more evenly distributed collection of documents over the six classes. A categorization was then performed by the SMO support vector classification built into the interface Weka under its built-in five-fold cross-validation examination (see Witten and Frank, 2005). We acchieve categorization results in the order of 40%-60%. Our results are correspond well to the results acchieved by others (Røst et al., 2006; Sun et al., 2002; Eichler, 2005).

As seen from Tables 4.4 and 4.5, sorted after performance, the linguistic preprocessing does indeed have a positive effect both on micro- and macro-level. What option performs best is difficult to judge as most render similar results. However, as pointed out, lemmatized text outperforms unlemmatized text. In lack of lemmatization, simple stop word removal helps to improve results.

Although only few documents belonging to the larger classes are mis-categorized into the smaller ones, they comprise a substantial part of the documents that gets categorized into the small ones. Macro-averaged precision is constantly lower than micro-averaged precision as the smaller classes generally have a poor precision. Thus the macro-averaged F1 is lower than the micro-averaged F1. The relatively poor macro-averaged F1 can thus be said to be a consequence of the uneven distribution of documents among the

Table 4.5: Macro-averaged support vector machine categorization.

Term type(s)	**Removal of functional words**	**Stop word removal**	**Macro-averaged precision**	**Macro-averaged recall**	**Macro-averaged** $F1$
Word forms	No	No	37.0%	44.5%	40.4%
Word forms	No	Yes	38.6%	47.8%	42.7%
Lemmata and word forms	No	No	42.1%	56.2%	48.2%
Lemmata and word forms	No	Yes	42.4%	56.1%	48.3%
Lemmata and word forms	Yes	No	40.8%	56.1%	47.2%
Lemmata and word forms	Yes	Yes	42.4%	57.1%	48.7%
Lemmata	No	No	43.4%	58.5%	**49.8%**
Lemmata	No	Yes	43.3%	56.8%	49.1%
Lemmata	Yes	No	42.1%	56.0%	48.1%
Lemmata	Yes	Yes	42.7%	58.3%	49.3%

classes. Recall is lowest for the smallest class and the two classes that lost a substantial part of their documents in the conversion from several classes per document to one.

There are no clear results patterns other than that lemmatization improves both precision and recall. It does so regardless whether the original word inflections are kept alongside or not. Stop word removal seems to take a middleground in performance—between unprocessed text and lemmatized text.

Chapter 5

Conclusion

After having examined a few categorization setups where documents have undergone different combinations of NLP preprocessing, the overall impression of the experiments is that NLP processing indeed can be utilized to improve the classification performance. The experiments does not however give an unanimous answer to what the ideal setup for categorization is, but does state that lemmatization, regardless of other implementational issues, is beneficial. Despite its simplicity, if lemmatization cannot be achieved, simple stop word removal has a clear positive impact on categorization.

5.1 Support Vector Machine Categorization

The effect of NLP preprocessing in case of SVM categorization is that it consistently increases categorization performance. This is tue to increase in both precision and recall. Lemmatization gives approximately 3% to 10% increase in F1 and a more even distribution of documents among the classes. This is reflected in the macro-averaged results, where the performance difference is larger.

5.2 String Matching Categorization

The results seem to be on par or better for the NLP preprocessed string matching categorization. In the analysis of how performance depends on the number of terms describing each class, we did not examine the previously better preprocessing combinations. Despite that, the two preprocessed combinations outperformed the not preprocessed setup. It is fair to assume that optimization fo the NLP preprocessing would yield yet higher difference.

The results are less clear for string matching than for SVM categorization. The general trend is that NLP preprocessing assists into better micro-averaged results (on average 3% better), but not necessarily much better macro-averaged (on average 1% better). The low macro-averaged increase might be an effect of the uneven document distribution in the corpus in our setup.

Using NLP tools, comparable performance can be achieved with fewer terms than whitout. Thus categorizing assisted by NLP preprocessing might be faster than without for certain systems.

The threshold, to where both precision and recall drops, is never crossed in the case when using lemmata only. This could mean that the system is more stable, but it could also be coincidental. Further investigation is needed, if conclusions are to be drawn.

5.3 Issues

The benefits achieved in these experiments are at times modest and might not motivate the additional computational cost of the part-of-speech tagging or the manual effort needed to implement and optimize a working system. But, if the implementation costs can be tolerated, and also time consuming preprocessing steps can be tolerated, or being performed ahead of time critical activities, such as human interaction, there clearly can be performance gains.

5.4 Comparison of results

Using a similar setup when categorizing emails in the closely related Swedish language Eichler (2005) concludes that stop word removal to have a negative effect. This is the opposite of our results. However, as she points out, in her case the negative effect is likely a consequence of the function oriented topical classes in the used setup.

Carlberger et al. (2001) conclude that stemming improve their information retrieval searches in Swedish by approximately 4%. This is in accordance to our lemmatization results. However, the results by Leopold and Kindermann (2002) from their SVM categorization of news articles in the more distant German language opposes this. They found lemmatization not to be beneficial due to the degree of inflectional forms.

Chapter 6

Further Work

6.1 Improved Classification Methods

The main classes are rather broad and subcategories in different branches may be overlapping. Classification onward by means of decision tree classification (Lewis and Ringuette, 1994) might form a more thorough test bed upon which the NLP tools can be tested. The setup would work as follows for the classification of a document:

- One (or more) main class would be chosen to be the topic best corresponding to the topic of the document.

- Within the chosen main class a new categorization would be conducted to choose one (or more) of its subcategories. The categorization would be performed in analogy with the choosing of main class, with the difference that only the documents belonging to the chosen main class and its subclasses would be used as document frequency corpus. Also there needs to be a threshold, below which the categorization would halt and no subclass would be chosen. This is to make sure the algorithm does not traverse the tree beyond what it is capable of classifying.

The categorization into the several different hierarchical levels calls for a more advanced evaluation metric since partial overlap in categorization will occur. The problems of our sparsly populated category setup remains. There might not be enough documents for a two level categorization in any branch of the hierarchical topic tree.

6.2 Multilingual Expansion

A relevant issue is the lack of manually classified documents in the right language, from which terms may be acquired. An alternative improvement would be to utilize more of the existing documents, either of mixed or different languages. However, this is bound to induce new kinds of errors. The introduction of non-Danish documents in a Danish classification, needs to be performed with careful hand and eye so as to maximize the benefits while suppressing errors. The author of this text is not aware of such a method.

6.3 Extracting Additional Terms

6.3.1 Complex Terms

Complex terms, consisting of more than one word, are generally more specific and less likely to be polysemous. Therefore they could serve categorization purposes better than single word terms. Terms involving more than one word often come in a variety of forms—shuffled, inflected and with other words blended in. Fully utilizing the potential of complex terms requires advanced tools to associate their different forms but without also incorporating similar construction with a different meaning. This also calls for a tuning process. Consult Jacquemin (2001) for a thorough overview.

A start would be to analyze the word collocations in the documents. The document collection is rather small, however, and that is an undesirable feature in word preference analysis. The benefits thus seem hazy. Some partial parsing might be introduced to acquire content relevant clusters. Multiword detection or noun group detection could also work in similar ways. Consult Nugues (2006) for more information on these subjects.

6.3.2 Decompounding

Danish grammar allows for a lot of compound words. Splitting compounds into their constituents could be valuable source for additional terms. Sjöbergh and Kann (2004) has investigated decompounding in the closely related Swedish language. For decompounding in information retrieval consult Karlgren (2005).

6.4 Classification Consistency

In a dynamic environment such as a collection of web links, an instance of a web page can accurately be seen as a sample or measurement of the web link in question. As the content of some sites varies over time, information may be accumulated, and events come and go. Classification is thus a challenge. The relevance of the topics associated to a site may vary. As has been documented in the literature, the consistency between indexers, or between the classifications of an indexer over time, is far from perfect (David and Giroux, 1995).

It is not clearly stated whether the hierarchical labels and keywords of the Arkade apply only to the referred sites themselves, or to the starting point for internet browsing as they represent—some documents seem to be of the former, and some of the latter. This insecurity affects the outcome of our setup as links between pages are not taken into account. The classification biases could be monitored by subject experts, who would compare the collected documents with the topics assigned to them by the Arkade classifiers.

6.5 Expanding the Corpus

Some documents in the collection are rather short, and instead of having to define a lower bound for the amount of text required to be a valid document, following the direct links from each page would mean that the content could be sufficiently expanded. The categorization would thus be based on a larger vocabulary and therefore less dependent on the presence of individual words, both in the single document, and the corpus as a whole. Hence, the results should become more stable.

This can either be done either for all documents or for documents shorter than a certain threshold. As most home pages contain weblinks, the corpus may be expanded not only to the direct links but to the links from the links and so on. As one proceeds onward one can expect to find that the information would tend to drift off from the topic of the original page, so there needs to be some sort of weighting based on distance from the initial document.

The language issue would have to be reconsidered as documents previously rejected as not being written in Danish may contain links to pages in Danish, and thus may be incorporated into our corpus, based on the text of the links. Previously deemed relevant sites will analogously contain irrelevant links. However, the number of documents would bo so large that manually examine the relevance of a document would not be feasible.

Chapter 7

Comments

7.1 Abbreviations

CMP – Classwise Macro-averaged Precision
CMR – Classwise Macro-averaged Recall
DMP – Document-wise Macro-averaged Precision
DMR – Document-wise Macro-averaged Recall
LTH – Lund Institute of Technology (sv.: Lunds Tekniska Högskola)
NLP – Natural Language Processing
POS tagger – Part-Of-Speech tagger
SMO – Sequential Minimal Optimization
SVM – Support Vector Machine
$tf \times idf$ – Term Frequency $\times$ Inverse Document Frequency
XML – eXtensible Markup Language

7.2 Comments to the Reader

If you have any opinion about what I'm doing please contact me at `f99je@efd.lth.se` All comments are appreciated.

/Jonas Ekedahl

Persons involved

Author

Jonas Ekedahl - Engineering Physics student at Lund Institute of Technology, in Lund, Sweden.

Supervisor

Pierre Nugues – Lecturer at the Computer Science Department, Lund Institute of Technology, in Lund, Sweden.

Co-supervisors

Birger Larsen - Associate Professor at the Department of Information Studies, Royal School of Library and Information Science, in Copenhagen, Denmark.

Anders Ardö - Associate Professor at the Department of Electrical and Information Technology, Lund Institute of Technology, in Lund, Sweden, as well as Author of the Combine Focused Crawler (Ardö and Golub, 2007).

Bibliography

Aizawa, A. (2001). Linguistic techniques to improve the performance of automatic text categorization. In *Proceedings of NLPRS-01, 6th Natural Language Processing Pacific Rim Symposium*, pages 307–314, Tokyo, JP.

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Inf. Process. Manage.*, 39(1):45–65.

Ardö, A. and Golub, K. (2007). Documentation for the combine (focused) crawling system. Web site: http://combine.it.lth.se/, Retrieved 6 February 2008.

Ardö, A. and Golub, K. (2008). Focused crawler software package. alvis project deliverable d7.2. Web site: http://www.it.lth.se/knowlib/publ/D7_2.pdf, Retrieved 6 February 2008.

Baeza-Yates, R. A. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*, chapter 2. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Bick, E. and Nygaard, L. (2007). Using Danish as a cg interlingua: A wide-coverage Norwegian-English machine translation system. In Joakim Nivre, Heiki-Jaan Kaalep, K. M. and Koit, M., editors, *Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007.*, pages 21–28.

Boser, B. E., Guyon, I., and Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In *Computational Learing Theory*, pages 144–152.

Brank, J., Grobelnik, M., Milic-Frayling, N., and Mladenic, D. (2002). Interaction of feature selection methods and linear classification models.

Carlberger, J., Dalianis, H., Hassel, M., and Knutsson, O. (2001). Improving precision in information retrieval for swedish using stemming. In *Proceedings of 13th Nordic Conference on Computational Linguistics NODALIDA 2001. Uppsala, Sweden*.

Corpus (2007). Word list of forms with information on part of speech and inflection. Danish Society for Language and Literature (DSL) Web site: http://korpus.dsl.dk/e-resurser/boejningsformer_ download.php?lang=en, Retrieved April 11, 2007.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Cristianini, N. and Shawe-Taylor, J. (2000). *An introduction to support Vector Machines: and other kernel-based learning methods*. Cambridge University Press, New York, NY, USA.

David, C. and Giroux, L. (1995). Indexing as problems solving: a cognitive appriach to consistency. In *Proceedings of CAIS/ACSI 95, 23rd Annual Conference of the Canadian Association for Information Science*, pages 78–89.

Eichler, K. (2005). Automatic classification of swedish email messages. Master's thesis, Eberhard-Karls-Universität, Tübingen.

Golub, K. (2006). Automated subject classification of textual web documents. *Journal of Documentation*, 62:350 – 371.

Golub, K. (2007). *Automated Subject Classification of Textual Documents in the Context of Web-Based Hierarchical Browsing*. PhD thesis, Department of Electrical and Information Technology, Lund University, Lund, Sweden.

Golub, K. and Ardö, A. (2005). Importance of html structural elements and metadata in automated subject classification. In *ECDL*, pages 368–378.

Henrichsen, P. J. (2007). A Norwegian letter-to-sound engine with Danish as a catalyst. In Joakim Nivre, Heiki-Jaan Kaalep, K. M. and Koit, M., editors, *Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007.*, pages 305–309.

Hisamitsu, T., Niwa, Y., Nishioka, S., Sakurai, H., Imaichi, O., Iwayama, M., and Takano, A. (1999). Term extraction using a new measure of term representativeness. In *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pages 475–484.

Jacquemin, C. (2001). *Spotting and Discovering Terms through Natural Language Processing*, volume 10. The MIT Press, New York, NY, USA.

Jacquemin, C. and Tzoukermann, E. (1997). NLP for term variant extraction: Synergy between morphology, lexicon and syntax.

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. In Nédellec, C. and Rouveirol, C., editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398, pages 137–142, Chemnitz, DE. Springer Verlag, Heidelberg, DE.

Karlgren, J. (2005). Compound terms and their constituent elements in information retrieval. In *15th Nordic Conference of Computational Linguistics*, Joensuu.

Keson, B. (2007). Vejledning til det danske morfosyntaktisk taggede parolekorpus. Danish Society for Language and Literature (DSL) Web site: http://korpus.dsl.dk/paroledoc_dk.pdf, Retrieved October 15, 2007.

Koller, D. and Sahami, M. (1997). Hierarchically classifying documents using very few words. In Fisher, D. H., editor, *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 170–178, Nashville, US. Morgan Kaufmann Publishers, San Francisco, US.

Leopold, E. and Kindermann, J. (2002). Text categorization with support vector machines. how to represent texts in input space? *Machine Learning*, 46(1-3):423–444.

Lewis, D. D. and Ringuette, M. (1994). A comparison of two learning algorithms for text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 81–93, Las Vegas, US.

Lewis, D. D., Yang, Y., Rose, T. G., and Li, F. (2004). RCV1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5:361–397.

Nugues, P. M. (2006). *An Introduction to Language Processing with Perl and Prolog. An Outline of Theories, Implementation, and Application with Special Consideration of English, French, and German.* Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Platt, J. C. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report, Redmond, Washington.

Porter, M. (2002). A Danish stop word list for snowball. Covered by the BSD License (see http://www.opensource.org/licenses/bsd-license.html), with Copyright (c) 2001, Dr Martin Porter, Web site: http://snowball.tartarus.org/algorithms/danish/stop.txt, Retrieved November 19, 2007.

Robertson, S. E. (1990). On term selection for queryexpansion. *Journal of Documentation*, 46(4):359–364.

Røst, T. B., Nytrø, Ø., and Grimsmo, A. (2006). Classifying encounter notes in the primary care patient record. In Stein, B. and Kao, O., editors, *Proceedings of the ECAI'06 3rd International Workshop on Text-based Information Retrieval TIR-06*.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523.

Sjöbergh, J. and Kann, V. (2004). Finding the correct interpretation of swedish compounds: A statistical approach. In *4th International Conference on Language Resources and Evaluation*.

Sun, A., Lim, E.-P., and Ng, W.-K. (2002). Web classification using support vector machine. In *WIDM '02: Proceedings of the 4th international workshop on Web information and data management*, pages 96–99, New York, NY, USA. ACM.

Sun, A., Lim, E.-P., and Ng, W.-K. (2003). Hierarchical text classification methods and their specification. In Chan, A. T., Chan, S. C., Leong, H. V., and Ng, V. T. Y., editors, *Cooperative Internet Computing*, pages 236–256. Kluwer Academic Publishers, Dordrecht, NL.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *HLT-NAACL*.

Toutanova, K. and Manning, C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora*, pages 63–70, Morristown, NJ, USA. Association for Computational Linguistics.

Weibel, S., Kunze, J., Lagoze, C., and Wolf, M. (1998). Dublin core metadata for resource discovery. Web site: http://portal.acm.org/citation.cfm?id=RFC2413#, Retrieved 6 February 2008.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, USA, 2 edition.

Appendix A

The Arkade Class Structure

The Arkade Class Structure as it was online on the 24th July 2007. Note that some classes share names.

A.1 How to Read the Table

In the fragment below *Billedkunst* 'Picture Art' is a main class that contains 14 documents. *Kunst i almindelighed* 'Art in general' is a subclass of *Billedkunst* that contains 53 documents. *Teori og æstetik* 'Theory and aesthetics' is a subclass of *Kunst i almindelighed* and contains 15 documents. *Fotografi* 'Photography' is a subclass of *Billedkunst* and contains eight documents.

1. Billedkunst (14)
    (a) Kunst i almindelighed (53)
        i. Teori og æstetik (15)
    (b) Fotografi (8)

A.2 The Class Structure

The complete class structure is given below. It consists of 323 classes whereof six are main classes.

1. Billedkunst (14)
    (a) Kunst i almindelighed (53)
        i. Teori og æstetik (15)
        ii. Kunsthistorie (12)
            A. Primitiv og forhistorisk kunst (1)
            B. Oldtidens kunst (7)
            C. Middelalderens kunst (13)
            D. Renæssancens kunst (4)
            E. Det 17. århundredes kunst (2)
            F. Det 19. århundredes kunst (5)
            G. Det 20. og 21. århundredes kunst (55)
        iii. Ikonografi, symbolik, allegori og emblematik (15)
        iv. Forfalskning (1)
        v. Lovgivning, ophavsret og kunststøtte (12)
        vi. Konkurrencer og priser (1)
        vii. Forskning og uddannelse (18)
        viii. Kunstforståelse og -formidling (31)
        ix. Organisationer, foreninger og kunsthandel (8)
            A. Organisationer (15)
            B. Kunstforeninger (2)
            C. Kunsthandlere (15)
        x. Museer, samlinger, gallerier (109)
    (b) Malerkunst (3)
        i. Historie
            A. Oldtidens malerkunst (1)
            B. Middelalderlig og byzantinsk malerkunst (5)
            C. Renæssancens malerkunst (1)
                - Det 15. århundrede (1)
            D. Det 19. århundredes malerkunst (4)
            E. Det 20. og 21. århundredes malerkunst (7)
        ii. Enkelte genrer og motiver (9)
        iii. Teknik og metode (3)
        iv. Museer, samlinger, gallerier (5)
        v. Malere A/Å (47)

(c) Tegnekunst (5)
    i. Historie (3)
        A. Middelalderlig og byzantinsk tegnekunst (4)
        B. Renæssancens tegnekunst (4)
            • Det 15. århundrede (2)
            • Det 16. århundrede (1)
        C. Det 17. århundredes tegnekunst (1)
        D. Det 18. århundredes tegnekunst (1)
        E. Det 19. århundredes tegnekunst (2)
        F. Det 20. og 21. århundredes tegnekunst (5)
        G. Ikke-vestlig tegnekunst (1)
    ii. Enkelte genrer og motiver (13)
    iii. Teknik og metode (3)
    iv. Museer, samlinger, gallerier (17)
    v. Tegnere A/Å (16)
(d) Grafik (34)
    i. Teori og æstetik (1)
    ii. Historie (9)
        A. Senmiddelalderens og renæssancens grafiske kunst (1)
            • Det 15. århundrede (2)
            • Det 16. århundrede (5)
        B. Det 17. århundredes kunst (4)
        C. Det 18. århundredes kunst (7)
        D. Det 19. århundredes kunst (7)
        E. Det 20. og 21. århundredes kunst (12)
        F. Ikke-vestlig grafik (7)
    iii. Enkelte genrer og motiver (37)
        A. Kunstnerbøger (18)
    iv. Teknik (10)
    v. Museer, samlinger, gallerier (28)
    vi. Grafikere A/Å (31)
(e) Skulptur (8)
    i. Historie (1)
        A. Primitiv og forhistorisk skulptur (1)

B. Oldtidens skulptur (4)
C. Middelalderlig og byzantinsk skulptur (6)
D. Renæssancens skulptur (1)
E. Det 19. århundredes skulptur (1)
F. Det 20. og 21. århundredes skulptur (2)
G. Ikke-vestlig skulptur (1)

ii. Enkelte genrer og motiver (4)
iii. Museer, samlinger, gallerier (6)
iv. Skulptører A/Å (22)

(f) Digital, multimedia og performancekunst (10)

i. Teori og æstetik (1)
ii. Digital kunst (11)
iii. Netkunst (69)
iv. Installationer (1)
v. Performance (3)
vi. Interaktiv kunst (7)
vii. Kropskunst / Body art (1)
viii. Lyskunst (1)
ix. Multimedia (3)
x. Museer, samlinger, gallerier (3)
xi. Kunstnere A/Å (5)

(g) Fotografi (8)

i. Teori og æstetik (7)
ii. Historie (4)
A. Det 19. århundrede (10)
B. Det 20. og 21. århundrede (1)
iii. Enkelte genrer og motiver (5)
iv. Fagfotografi (1)
v. Teknik og metode (3)
A. Historiske teknikker (6)
- Daguerreotypi (3)

B. Firmaer (2)
vi. Billedkilder (5)
vii. Forskning og uddannelse (2) Organisationer og foreninger (1)
A. Foreninger (1)

viii. Museer, samlinger, gallerier (15)
ix. Fotografer A/Å (16)

2. Arkitektur, planlægning og byggeteknik (46)

(a) Arkitektur, bygningsplanlægning (45)

i. Teori og æstetik (28)
ii. Historie (18)
A. Oldtidens arkitektur (6)
B. Middelalderens arkitektur (10)
C. Det 17. århundredes arkitektur (1)
D. Det 18. århundredes arkitektur (1)
E. Det 19. århundredes arkitektur (4)
F. Det 20. og 21. århundredes arkitektur (11)
iii. Politik og administration (2)
iv. Teknik og metode (11)
A. Industrialiseret byggeri (1)
B. Traditionelt byggeri (2)
C. Økologisk og bæredygtigt byggeri (4)
D. Indeklima, arbejdsmiljø (2)
• Lys, akustik, varme og ventilation (3)
E. Tilgængelighed (7)
v. Forskning og uddannelse (45)
vi. Konkurrencer og priser (6)
vii. Arkitektvirksomhed (6)
viii. Organisationer og foreninger (17)
ix. Bygninger og projekter (24)
A. Religiøse bygninger (9)
B. Sundhedsinstitutioner (1)
C. Uddannelses- og forskningsinstitutioner (4)
D. Kulturinstitutioner (8)
E. Fremstillingsvirksomhed (1)
F. Administration (1)
G. Landbrug (1)
H. Boliger (11)
x. Tegnestuer og arkitekter A/Å (102)

xi. Museer, samlinger, gallerier (5)

(b) Planlægning (12)

i. Planlægning i almindelighed

A. Teori og æstetik (2)

B. Historie (5)

C. Politik og administration (1)

D. Teknik og metode (10)

- Byøkologi (5)

E. Forskning og uddannelse (7)

F. Konsulentvirksomhed (2)

ii. Arealplanlægning (3)

A. Teori og æstetik (1)

B. Politik og administration (1)

C. Plantyper (3)

- Landsplanlægning (2)
- Regionplanlægning (7)
- Kommuneplanlægning (2)
- Lokalplanlægning (4)

iii. Landskabsplanlægning (4)

A. Teori og æstetik (1)

B. Historie (2)

C. Teknik og metode (1)

D. Plantyper (1)

iv. Byplanlægning (25)

A. Teori og æstetik (8)

B. Historie (6)

C. Politik og administration (2)

D. Teknik og metode (5)

E. Plantyper (4)

- Bydelsplanlægning (1)
- Landsbyplanlægning (1)

(c) By- og landskabsarkitektur, havekunst (7)

i. Teori og æstetik (2)

ii. Historie (1)

iii. Plantyper (1)

(d) Byfornyelse, bygningsbevaring og restaurering (11)

i. Byer og bydele (13)

ii. Bygninger (9)

(e) Byggeteknik (27)

3. Kunsthåndværk og design (17)

(a) Kunsthåndværk og design i almindelighed (69)

i. Teori, æstetik og formlære (11)

ii. Historie (4)

iii. Designmetoder og -processer (18)

iv. Designmanagement (3)

v. Lovgivning og kunststøtte (3)

vi. Tilgængelighed - design til ældre, handicappede, børn etc. (8)

vii. Økologiske design (9)

viii. Konkurrencer og priser (2)

ix. Forskning og uddannelse (54)

x. Foreninger og handel med kunsthåndværk (27)

xi. Organisationer (15)

xii. Museer, samlinger, gallerier (42)

xiii. Kunsthåndværkere og designere A/Å (116)

(b) Materialer (5)

i. Teori og æstetik (2)

ii. Metaller (11)

iii. Læder, skind, ben og horn (2)

iv. Plastik (1)

v. Træ (2)

vi. Andre materialer (5)

(c) Møbler (33)

i. Enkelte møbeltyper (4)

ii. Møbler til særlig anvendelse (1)

(d) Keramik (47)

i. Teori og æstetik (3)

ii. Historie (1)

iii. Teknik og metode (3)

iv. Keramik til særlig anvendelse (2)

(e) Tekstiler (28)

i. Teori og æstetik (3)
ii. Historie (5)
iii. Teknik og metode (5)
iv. Enkelte typer af tekstiler (9)
v. Tekstiler til særligt brug (5)

(f) Glas (23)

i. Teori og æstetik (1)
ii. Historie (2)
iii. Teknik og metode (1)
iv. Glasmaleri (2)

(g) Smykker (43)

(h) Dragter og mode (18)

i. Teori og æstetik (3)
ii. Historie (11)
iii. Teknik og metode (1)
iv. Beklædning til særlig anvendelse (5)
v. Tilbehør (3)

(i) Industriel design (17)

i. Teknik og metode (2)
ii. Enkelte produkttyper (3)

(j) Visuel kommunikationsdesign (8)

i. Teori og æstetik (5)
ii. Historie (5)
iii. Teknik og metode (1)
iv. Anvendelsesområder (4)
A. Plakater (8)
B. Logoer og skrift (4)
v. Digital design (11)
A. Web-design (4)
B. Interfacedesign (2)

(k) Rumdesign (1)

i. Teori (2)
ii. Historie (2)
iii. Boligindretning (1)
iv. Skoler, museer og anden institutionsindretning (2)
v. Hotel- og restaurationsindretning (1)
vi. Kontor- og arbejdspladsindretning (1)
vii. Rumkunst (2)
viii. Lys (1)
ix. Akustik (1)

(l) Scenografi (1)
i. Teater (1)
A. Teaterdekorationer og kostumetegninger (3)
B. Teaterhistorie (5)
C. Teaterteknik (1)
ii. Animation (1)

4. Museologi (11)

(a) Teori (19)
(b) Historie (13)
(c) Etik (8)
(d) Pædagogik (16)
(e) Museografi (7)
i. Lovgivning (2)
ii. Museumsbyggeri, -drift og -sikkerhed (13)
iii. Indsamling (6)
iv. Genstandsregistrering (7)
v. Udstillingsteknik (6)
vi. Museums- og udstillingskataloger (1)
vii. Museumsuddannelse (8)
(f) Museumspublikum (13)
(g) Organisationer og foreninger (1)
i. Internationale organisationer (9)
ii. Nationale organisationer (12)
iii. Samarbejdsprojekter (11)

5. Konservering (23)
    (a) Historie og teori (3)
    (b) Teknisk museologi (5)
        i. Biologiske nedbrydning (5)
        ii. Lys, luft, klima (4)
        iii. Udstillingsteknik, magasin, sikring mm. (1)
    (c) Teknik og metode (9)
        i. Analysemetoder (7)
        ii. Maleteknikker (3)
        iii. Feltkonservering (3)
        iv. Grafiske trykteknikker (4)
        v. Digital billedbehandling (1)
    (d) Dokumentation (2)
        i. Fotografisk dokumentation (3)
    (e) Kemi
        i. Farver (2)
    (f) Materialer
        i. Grafisk materiale (9)
        ii. Kulturhistoriske genstande (14)
        iii. Kunst (3)
        iv. Monumentalkunst (12)
        v. Naturhistoriske genstande (1)
    (g) Naturbevaring (2)
    (h) Lovgivning, standarder, code of ethics mm. (5)
    (i) Forskning og uddannelse (9)
    (j) Organisationer og foreninger (18)
6. Kunstnere, fotografer, arkitekter, designere A/Å (36)
    (a) A (23)
    (b) B (35)
    (c) C (19)
    (d) D (16)

(e) E (9)
(f) F (16)
(g) G (26)
(h) H (33)
(i) I (3)
(j) J (18)
(k) K (24)
(l) L (21)
(m) M (25)
(n) N (8)
(o) O (3)
(p) P (26)
(q) R (15)
(r) S (35)
(s) T (13)
(t) V (6)
(u) W (8)
(v) Y (1)
(w) Ø (1)
(x) Å (1)

Appendix B

Example Documents

B.1 A Document Belonging to One Class

This is an example document that belongs to only to the main class *Billedkunst* 'Picture art'. The full hierarchical classification would be *Billedkunst - Grafik - Historie - Det 20. og 21. århundredes kunst* 'Picture art - Graphics - History - The art of the 20th and 21st centuries'.

```
<?xml version="1.0" encoding="UTF-8"?>
 <documentRecord id="0039666E6B2F880C317535BA119C0005">
  <acquisition>
   <acquisitionData>
    <modifiedDate>2006-10-29 17:10:35</modifiedDate>
    <checkedDate>2006-11-02 13:22:08</checkedDate>
    <httpServer>Microsoft-IIS/6.0</httpServer>
    <urls>
     <url>http://www.roedemor.dk/</url>
    </urls>
   </acquisitionData>
  <canonicalDocument>
   <section> RØDE MOR BOG, RØDE MOR ROCKSHOW, RØDE MOR
             CDboxsæt, RØDE MOR grafik UDSTILLING mm, Troels
             Trier, Peter Ingemann, Henrik Strube, Lars Trier,
             Erik Clausen , Leif Sylvester Petersen, RØDE MOR
             gruppen RØDE MOR (1969-1978) Det progressive,
             samfunds-satiriske og politisk velfunderede
             kunstnerkollektiv RØDE MOR har genudgivet en stor
             del af deres omfattende medie-produktion. RØDE MOR
             bestod af FORFATTERE, SATIRIKERE, KUNSTMALERE,
```

GRAFIKERE, MUSIKERE og multimedieshowet RØDE MOR's
ROCKCIRKUS. RØDE MOR arrangerede utallige
kunstudstillinger og udgav tonsvis af materiale -
musik - bøger - kataloger - billeder -pamfletter -
sangbøger, incl et årligt manifest, der beskrev,
hvordan det stod til på den politiske scene og med
dem selv. Det første manifest blev udgivet i 1970.
RØDE MOR indstillede den kunstneriske produktion i
1978 og oprettede i stedet en fond af indkomsten
fra salg af plader og plakater. Fonden støttede
politisk kunst på den danske venstrefløj MEDLEMMER
AF RØDE MOR: Dea Trier Mørch - Forfatter og
Billedkunstner Troels Trier - Billedkunstner,
Skuespiller, Sanger og Sangskriver Andreas Trier
Mørch - Fotograf og Arkitekt Lars Trier -
Guitarist, Komponist Jacob Trier Peter Ingemann -
Musiker, Komponist og Revisor Leif Sylvester
Petersen - Billedkunstner, Skuespiller, Sanger og
Sangskriver Erik Clausen - Billedkunstner,
Filminstruktør, Sangskriver, Skuespiller Ole Thilo
- Musiker, Komponist Henrik Strube - Guitarist,
Sanger, Sangskriver, Story-entertainer Ann
Thorsted Finn Sørensen Karsten Sommer - Studiemand
John Ravn - Trommespiller Michael Boesen -
Fotograf, Guitarist, Skuespiller Niels Brunse -
Forfatter Alice Faber - Billedkunstner Poul
Poulsen - Musiker Dorte Fasting Ole Finding -
Billedkunstner Tommy Flugt - Forfatter og
Billedkunstner Pia Funder Per Almar Johnson,
Billedkunstner Maiken Junker Anne-Mette Kruse
Thomas Kruse - Billedkunstner Erling Benner Larsen
Kim Menzer - Musiker Peter Mogensen - Musiker
Yukari Ochiai Jens Asbjørn Olesen Anne- Marie
Steen Petersen - Forfatter og Billedkunstner
</section>
</canonicalDocument>
<metaData>
<meta name="title">RØDE MOR BOG, RØDE MOR ROCKSHOW, RØDE MOR
CDboxsæt, RØDE MOR grafik UDSTILLING mm, Troels Trier,
Peter Ingemann, Henrik Strube, Lars Trier, Erik
Clausen , Leif Sylvester Petersen, RØDE MOR gruppen

```
      </meta>
 <meta name="dc:format">text/html</meta>
 <meta name="dc:format">text/html; charset=iso-8859-1</meta>
 <meta name="dc:creator">work 'n' web Helle Fastrup</meta>
 <meta name="dc:description">Røde Mor gruppen's kulturkritik
      er stadig aktuel i dagens Danmark. Med
      Danmarkshistoriens smukkeste CD boxsæt med indlagt bog
      og grafik om og af Røde Mor, har  Røde Mor indskrevet
      sig i det nye årtusinde.</meta>
 <meta name="dc:subject"> RØDE MOR </meta>
 <meta name="dc:subject">rødemor ROCKSHOW</meta>
 <meta name="dc:subject">Røde Mor CDboxsæt </meta>
 <meta name="dc:subject">rødemor boxsæt </meta>
 <meta name="dc:subject">Røde Mor's Rockcirkus </meta>
 <meta name="dc:subject">rødemorboxsæt </meta>
 <meta name="dc:subject">røde mor </meta>
 <meta name="dc:subject">Røde Mor</meta>
 <meta name="dc:subject">Troels Trier </meta>
 <meta name="dc:subject">Dea Trier Mørch </meta>
 <meta name="dc:subject">Trier </meta>
 <meta name="dc:subject">satire </meta>
 <meta name="dc:subject">RØDE MOR's Rockcirkus </meta>
 <meta name="dc:subject">Kunstnerkollektiver </meta>
 <meta name="dc:subject">Erik Clausen </meta>
 <meta name="dc:subject">Peter Ingemann </meta>
 <meta name="dc:subject">Leif Sylvester Petersen </meta>
 <meta name="dc:subject">Andreas Trier Mørch </meta>
 <meta name="dc:subject">Lars Trier </meta>
 <meta name="dc:subject">Jacob Trier </meta>
 <meta name="dc:subject">Proletarisk kunst </meta>
 <meta name="dc:subject">politisk rockscene </meta>
 <meta name="dc:subject">internationale </meta>
 <meta name="dc:subject">hjemlig hygge </meta>
 <meta name="dc:subject">Rok Ork </meta>
 <meta name="dc:subject">illustreret sangbog</meta>
 <meta name="dc:subject">illustrede sangbøger</meta>
</metaData>
<links>
 <outlinks>
  <link type="frame">
   <location>http://www.roedemor.dk/left-top.htm</location>
```

```
    </link>
    <link type="frame">
     <location>http://www.roedemor.dk/top.htm</location>
    </link>
    <link type="frame">
     <location>http://www.roedemor.dk/left.htm</location>
    </link>
    <link type="frame">
     <location>http://www.roedemor.dk/main.htm</location>
    </link>
    <link type="frame">
     <location>http://www.roedemor.dk/right.htm</location>
    </link>
    <link type="frame">
     <location>http://www.roedemor.dk/UntitledFrame-1.htm
               </location>
    </link>
    <link type="frame">
     <location>http://www.roedemor.dk/UntitledFrame-2.htm
               </location>
    </link>
   </outlinks>
  </links>
  <analysis>
   <property name="language">da</property>
  </analysis>
 </acquisition>
</documentRecord>
```

B.2 A Document Belonging to Three Classes

This sample document belongs to the three main classes *Billedkunst* 'Picture art', *Kunsthåndværk og design* 'Crafts and design' and *Kunstnere, fotografer, arkitekter, designere A-Å* 'Artists, photographers, architects and designers A-Z'.

In the full hierarchy the documents belongs to five classes:

- *Billedkunst - Malerkunst - Historie - Det 20. og 21. århundredes malerkunst*

  'Picture art - Art painting - History - The art painting of the 20th and

21st centuries'

- *Billedkunst - Malerkunst - Malere A-Å*

  'Picture art - Art painting - Painters A-Z'

- *Kunsthåndværk og design - Kunsthåndværk og design i almindelighed - Kunsthåndværkere og designere A-Å*

  'Crafts and design - Crafts and design in general - Handcrafters and designers A-Z'

- *Kunsthåndværk og design - Dragter og mode*

  'Crafts and design - Clothes and fashion'

- *Kunstnere, fotografer, arkitekter, designere A-Å - L*

  'Artists, photographers, architects and designers A-Z - L'

```
<?xml version="1.0" encoding="UTF-8"?>
<documentRecord id="36D54468579782565A566A3FE2EA4885">
 <acquisition>
  <acquisitionData>
   <modifiedDate>2005-05-26 21:00:10</modifiedDate>
   <checkedDate>2006-11-02 12:39:57</checkedDate>
   <httpServer>Apache/1.3.37 (Unix) Sun-ONE-ASP/4.0.2 PHP/
               4.4.4 m</httpServer>
   <urls>
    <url>http://www.zahidlatif.com/</url>
   </urls>
  </acquisitionData>
  <canonicalDocument>
   <section> Zahid Latif Art  IMG   IMG   IMG   IMG
             english 1024 X 768 pixels IE 4.0 + for best
             result updated 1 june 2005 </section>
  </canonicalDocument>
  <metaData>
   <meta name="title">Zahid Latif Art</meta>
   <meta name="dc:format">text/html</meta>
   <meta name="dc:format">text/html;charset=iso-8859-1</meta>
   <meta name="dc:description">http//:uv.dk-designskole.dk/
         di4p/zahid studerende på Danmarks Designskole,
         præsentation af sine værker: akvarel, akryl,
         kobbertryk, kalligrafi, tegninger, skoleprojekter,
```

```
        herunder slideshow fra Lahore osv. Æstetisk
        webprodukt med focus på design og form. </meta>
  </metaData>
  <links>
   <outlinks>
    <link type="img">
     <location>http://www.zahidlatif.com/pict/hviidkasse.gif
               </location>
    </link>
    <link type="a">
     <location>http://www.zahidlatif.com/home.html</location>
    </link>
   </outlinks>
  </links>
  <analysis>
   <property name="language">da</property>
  </analysis>
 </acquisition>
</documentRecord>
```

Appendix C

Stop Words

A Danish stop word list. Comments begin with a dash. Each stop word is at the start of a line. The list is ranked so that the generally most common word comes first and the rarest is last. For more information consult Porter (2002).

og	—	and
i	—	in
jeg	—	I
det	—	that (dem. pronoun)/it (pers. pronoun)
at	—	that (in front of a sentence)/to (with infinitive)
en	—	a/an
den	—	it (pers. pronoun)/that (dem. pronoun)
til	—	to/at/for/until/against/by/of/into, more
er	—	present tense of "to be"
som	—	who, as
på	—	on/upon/in/on/at/to/after/of/with/for, on
de	—	they
med	—	with/by/in, along
han	—	he
af	—	of/by/from/off/for/in/with/on, off
for	—	at/for/to/from/by/of/ago, in front/before, because
ikke	—	not
der	—	who/which, there/those
var	—	past tense of "to be"
mig	—	me/myself
sig	—	oneself/himself/herself/itself/themselves

men — but
et — a/an/one, one (number), someone/somebody/one
har — present tense of "to have"
om — round/about/for/in/a, about/around/down, if
vi — we
min — my
havde — past tense of "to have"
ham — him
hun — she
nu — now
over — over/above/across/by/beyond/past/on/about, over/past
da — then, when/as/since
fra — from/off/since, off, since
du — you
ud — out
sin — his/her/its/one's
dem — them
os — us/ourselves
op — up
man — you/one
hans — his
hvor — where
eller — or
hvad — what
skal — must/shall etc.
selv — myself/youself/herself/ourselves etc., even
her — here
alle — all/everyone/everybody etc.
vil — will (verb)
blev — past tense of "to stay/to remain/to get/to become"
kunne — could
ind — in
når — when
være — present tense of "to be"
dog — however/yet/after all
noget — something
ville — would
jo — you know/you see (adv), yes

deres — their/theirs
efter — after/behind/according to/for/by/from, later/afterwards
ned — down
skulle — should
denne — this
end — than
dette — this
mit — my/mine
også — also
under — under/beneath/below/during, below/underneath
have — have
dig — you
anden — other
hende — her
mine — my
alt — everything
meget — much/very, plenty of
sit — his, her, its, one's
sine — his, her, its, one's
vor — our
mod — against
disse — these
hvis — if
din — your/yours
nogle — some
hos — by/at
blive — be/become
mange — many
ad — by/through
bliver — present tense of "to be/to become"
hendes — her/hers
været — be
thi — for (conj)
jer — you
sådan — such, like this/like that

Appendix D

The PAROLE Word Class Tags

The word class tag set we have trained the part-of-speech tagger to tag into is the PAROLE Danish tag set. These tags consists of alphanumeric codes. The first letter is the word class. The following numbers and letters are codes for different attributes specific for the word class. The codes are arranged in a positional notation, where each position referres to a specific attribute. For more information consult Keson (2007).

An example of what a tagged text might look like:
Hvem/PT-C[SP]U–U har/VADR=—-A- lavet/VAPA=S[CN]I[ARU]-U bogen/NCCSU==D ?/XP Det/PP3NSU-NU har/VADR=—-A- digteren/NCCSU==D ./XP

CatGram	Attribute	Value	Tag	Position
Adjective			A	1
	SsCatGram	Cardinal	C	2
		Normal	N	2
		Ordinal	O	2
	Degree	Positive	P	3
		Comparative	C	3
		Superlative	S	3
		Absolute Superl.	A	3
	Gender	Common	C	4
		Neuter	N	4
	Number	Singular	S	5
		Plural	P	5
	Case	Unmarked	U	6
		Genitive	G	6
	Definiteness	Definite	D	8
		Indefinite	I	8
	Use	Adverbial Use	R	9
		Unmarked	U	9
Adposition			S	1
	SsCatGram	Preposition	P	2
Adverb			R	1
	SsCatGram	General	G	2
	Degree	Positive	P	3
	Comparative		C	3
	Superlative		S	3
	Absolute Superl.		A	3
	Unmarked		U	3
Conjunction			C	1
	SsCatGram	Coordinative	C	2
		Subordinative	S	2
Interjection			I	1

CatGram	Attribute	Value	Tag	Position
Noun			N	1
	SsCatGram	Proper	P	2
		Common	C	2
	Gender	Common	C	3
		Neuter	N	3
	Number	Singular	S	4
		Plural	P	4
	Case	Unmarked	U	5
		Genitive	G	5
	Definiteness	Definite	D	8
		Indefinite	I	8
Pronoun			P	1
	SsCatGram	Personal	P	2
		Demonstrative	D	2
		Indefinite	I	2
		Interrog./relative	T	2
		Reciprocal	C	2
		Possessive	O	2
	Person	First	1	3
		Second	2	3
		Third	3	3
	Gender	Common	C	4
		Neuter	N	4
	Number	Singular	S	5
		Plural	P	5
	Case	Nominative	N	6
		Genitive	G	6
		Unmarked	U	6
	Possessor	Singular	S	7
		Plural	P	7
	Reflexive	Yes	Y	8
		No	N	8
	Register	Formal	F	9
		Obsolete	O	9
		Polite	P	9
		Unmarked	U	9

CatGram	Attribute	Value	Tag	Position
Residual			X	1
	SsCatGram	Abbreviation	A	2
		Foreign Word	F	2
		Punctuation	P	2
		Formulae	R	2
		Symbol	S	2
		Other	X	2
Unique			U	1
Verb			V	1
	SsCatGram	Main	A	2
		Medial	E	2
	Mood	Indicative	D	3
		Imperative	M	3
		Infinitive	F	3
		Gerund	G	3
		Participle	P	3
	Tense	Present	R	4
		Past	A	4
	Number	Singular	S	6
		Plural	P	6
	Gender	Common	C	7
		Neuter	N	7
	Definiteness	Definite	D	8
		Indefinite	I	8
	Use	Adjectival Use	A	9
		Adverbial Use	R	9
		Unmarked	U	9

Appendix E

The Flexion Word Class Tags

Our dictionary is based on the word list Flexion, that uses a set of one letter tags to indicate the word classes. An additional number specifies inflectional form. The numbers denote different inflections depending on main word class. The inflection tags are not used in this work as we are interrested only in the main word class. For further information on Flexion see Corpus (2007).

The different word class tags used in Flexion:

Tag	**Word Class (English)**	**Word Class (Danish)**
A	Adjective	Adjektiv
Æ	Preposition	Præposition
D	Adverb	Adverbium
F	Abbreviation	Forkortelse
I	Prefix	Præfiks
K	Conjunction	Konjunktion
L	Onomatopoeia	Lydord
O	Pronoun	Pronomen
P	Proper noun	Proprium
S	Noun	Substantiv
T	Number	Talord
U	Interjection	Udråbsord
V	Verb	Verbum
X	Unidentified	Uidentificeret

An example of an entry in Flexion. This is the entry for the noun *certifikat*, with its eight inflectional forms:

```
*
certifikat
S
2 certifikat
4 certifikats
8 certifikatet
16 certifikatets
32 certifikater
64 certifikaters
128 certifikaterne
256 certifikaternes
```