Covid-19 related Text Mining in Medical Articles

Using Scispacy with pretrained models

TABLE OF CONTENTS

1 Introduction

2 Method

3 Results

4 Improvements

Introduction - Project

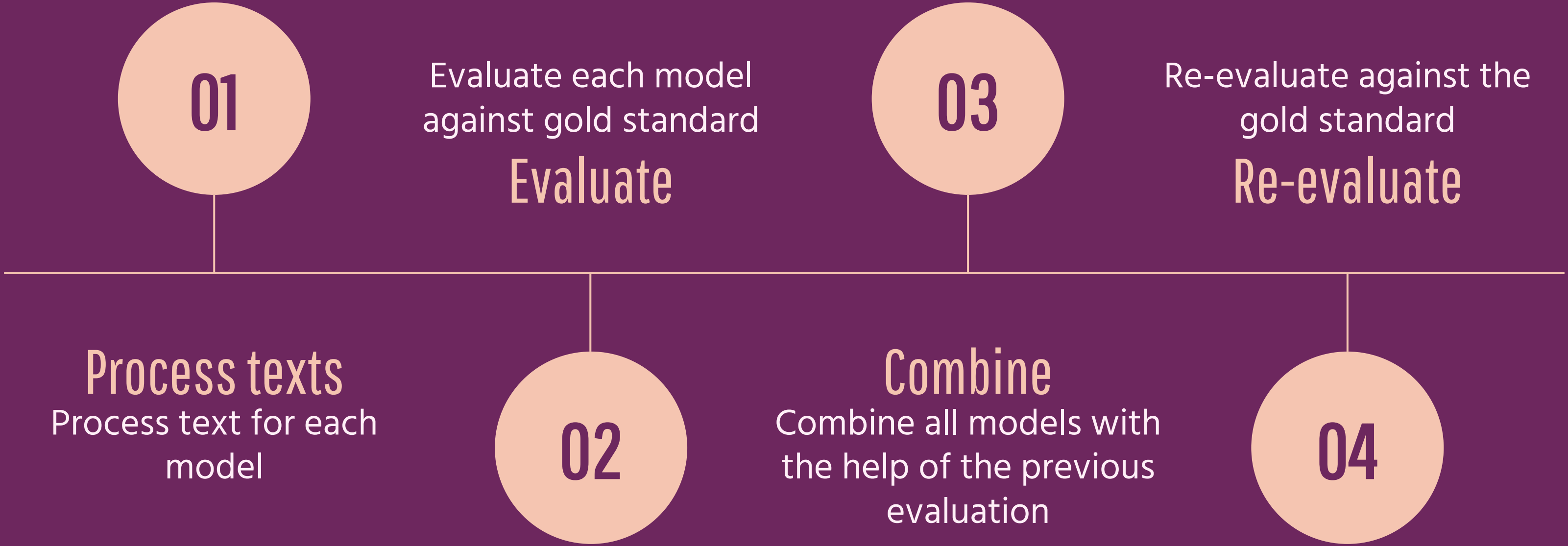
The project is done in cooperation with Sonja Aits, Pierre Nugues and other groups from the course

- Contribute to research of Covid-19
- Why use text mining?
- Using Named-Entity-Recognition (NER)
- Scispacy: Pretrained NER models

Introduction - Scispacy

Model	Protein	Species	Chemical	Disease
Bionlpg13cg	✓	✓	✓	✓
Craft	✓	✓	✓	
Bc5cdr			✓	✓
Jnlpba	✓			

Method



01

Process texts

Process text for each model

02

Evaluate

Evaluate each model against gold standard

03

Combine

Combine all models with the help of the previous evaluation

04

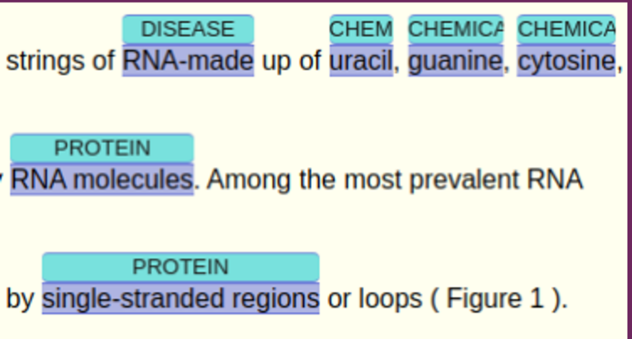
Re-evaluate

Re-evaluate against the gold standard

01

Process Texts

- Texts from Kaggle Open Research
- JSON format
- Load models and texts
- Apply models for each text



strings of RNA-made up of uracil, guanine, cytosine,

RNA molecules. Among the most prevalent RNA

by single-stranded regions or loops (Figure 1).

02

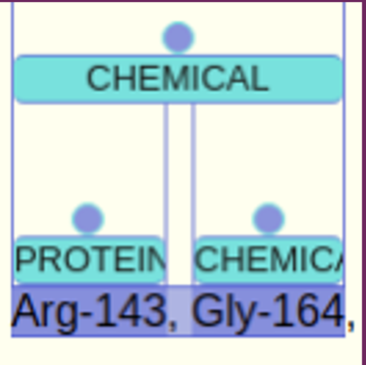
Evaluate

- Use processed texts
- Compare with gold standard
- Evaluation script

03

Combine

- Combine output of models
- Length conflict -> Longest match
- Class conflict -> evaluation scores
- Edge case -> Iterative removal



04

Re-evaluate

- Use combined output
- Compare with gold standard

Results - Harmonic Mean (%)

Model	Protein	Species	Chemical	Disease	All Classes
Bionlpg13cg	59	12		0	24
Craft	64	18			41
Bc5cdr				33	33
Jnlpba	56				56
Combined	44	14		32	31

Improvements

- Evaluate more classes
- Evaluate on appropriate standard
- Smarter solution than longest match
- Don’t combine - Use best model

THANKS

Do you have any questions?

William Lindholm
dat15wli@student.lu.se

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**