Examensarbete:

Modeller och arkitektur för n-gram-baserad språkhjälp.

**Din arbetsplats**

Oribi utvecklar datorstöd för personer med läs- och skrivsvårigheter. Oribi ligger bl. a. bakom rättstavningsprogrammen ”Stava Rex” för svenska, ”SpellRight” för svenskar som skriver på engelska, ordprediktionsprogrammet Saida och SAOL Plus, som är en datorbaserad version av Svenska akademiens ordbok med en uppslagsfunktion som även kan hantera felstavningar. Oribis sökmetod hittar det avsedda ordet. Den härleder *tjosk* till *kiosk*, *asusigera* till *associera* och *skunalist* till *journalist*. Sökmetoden kan ses i arbete på Nationalencyklopediens webbsajt.

**Din uppgift**

N-gram används för att göra sannolikhetsmodeller av språk, och är baserade på stora textmassor. Googles n-gram-korpus tillhandahåller n-gram upp till n=5. Med den som bas kan man hitta felstavade ord som av misstag har skrivits som andra ord, s.k. ”real word errors”. En så stor databas som n-gram-korpusen kan enbart hanteras av en eller helst fler servrar, det är uteslutet att idag hantera hela databasen i ett program lokalt på användarnas datorer. Följande autentiska sekvenser innehåller fel som endast kan upptäckas med hjälp av kontexten.

Jag har släkt lampan.

I fell like I'm being set up for a bad day.

Uppgiften kan vara en lämplig kombination av följande:

* Konstruera och utvärdera en bra och enkel sannolikhetsmodell för att upptäcka fel givet n-gramdata.
* Konstruera och utvärdera en datastruktur och en skalbar serverbaserad arkitektur för att hantera förfrågningar från en rättstavningsklient.

**Dina kvalifikationer**

Examensarbetet lämpar sig för en datalingvist/språkteknolog eller en teknolog med språkintresse. Kunskaper i grundläggande sannolikhetsteori och programmeringsvana är nödvändiga.

**Till ditt stöd**

får du Anders Holtsberg och Caroline Willners på Oribi, samt Pierre Nugues på Institutionen för datavetenskap, Lunds Tekniska Högskola.

|  |  |  |
| --- | --- | --- |
| **Placering**  Oribi, Kungsgatan 2c, Lund | **Kontaktpersoner**  Caroline Willners  Oribi AB  046-32 30 12  Caroline.Willners@oribi.se | Pierre Nugues  Datavetenskap, LTH  046-222 96 40  Pierre.Nugues@cs.lth.se |