Text Mining
Introduction

Mattias Villani

Division of Statistics
Dept. of Computer and Information Science
Linköping University

What is Text Mining?

- **Quantitative analysis** of natural language **texts**.
- Combination of **statistics**, **linguistics** and **computer science**.
- Stages:
  - **Reading and organizing textual data** in suitable computer format
  - Understanding the **linguistic structure** of the data
  - **Probabilistic models** of the language

Some examples of text mining

- **Language models** (predict the next word on smartphone keyboard)
- **Machine translation** (Google translate)
- **Document classification** (did Shakespeare write this sonnet? Spam and blog filters)
- **Information retrival** (Google search)
- **Sentiment analysis** (positive/negative sentiment in tweets)
- **Part-of-speech tagging** (classify the grammatical category of words in a sentence)

Intersection of three subjects

- **Databases** and **information retrival**
  - Professor Patrick Lambrix, ADIT
- **Computational linguistics**
  - Docent Marco Kuhlmann, CiltLab
- **Statistics**
  - Mattias Villani, professor Statistics
  - Oleg Sysoev, lecturer Statistics

  Måns Magnusson, PhD student Statistics

- [Fine print: This course is a bold cross-disciplinary experiment!]

Course outline

- **Introductory modules** (pick at least 2 out of 3):
  - Introduction to Python programming Johan Falkenjack
  - Introduction to statistical modeling (Sysoev)
  - Introduction to computational linguistics Kuhlmann
- **Data models and Information Retrieval for Textual Data** Lambrix
- **Statistical Models for Textual Data** Magnusson
- **Text Mining Project** (Villani and friends)

Examination

- **Computer labs, 3 credits**
  - Should be performed in **pairs of students**
  - Graded Pass/Fail.
- **Text mining project, 3 credits**
  - **Individual**
  - Graded on the ECTS scale (A-F)
  - Concisely **written project report**
  - **Oral presentation** on Jan 22
  - Student should come up with their own ideas for the project. Project proposals should be sent to Mattias Villani no later than November 20.
- **Ph.D. students** are required to do a more ambitious text mining project. Grades for Ph.D. students are Pass/Fail.

Course literature

- The internet ...
- **Natural Language Processing with Python**. Contains a lot of practical hands-on material using the NLTK toolkit for Python.
- **Foundations of Statistical Natural Language Processing**. Contains the background theory for computational linguistics and statistical analysis of text data.
- Extra material.
- Both books are free (gratis) in electronic versions, see the course webpage. The books have not been ordered to the campus bookstores.

Computing

- The computers in the lab room **Statistik PUL** (right across the room of Peter Nilsson in the E-building) have:
  - **Python(x,y)** (Python + IDE)
  - **NLTK toolkit**
  - **RStudio** (R + IDE)
- On a Linux or Mac:
  - **Spyder IDE** for Python
  - Scientific packages **numpy** and **scipy** (in the repositories)
  - Plotting module **matplotlib**
- Installing NLTK: http://nltk.org/install.html
- You may also want to install the python modules:
  - **beautiful soup** (for reading web pages)
  - **twitter** (access to Twitter's API from Python).
- **Text mining packages in R**, see the the **tm** package and http://cran.r-project.org/web/views/NaturalLanguageProcessing.html

Live demo of NLTK and Python

- Getting started with NLTK.
- Movie review example.