807 - TEXT ANALYTICS
http://csee.essex.ac.uk/staff/poesio/Teach/807/

Massimo Poesio
poesio@essex.ac.uk

Lecture 1: Introduction, Preprocessing

Outline of lecture

• **Text analytics as a solution to the big data problem**
• Contents of the module
• Preprocessing
• Existing tools
• The role of statistics in NLP
• Module logistics
The Big Data Challenge

- We have produced more data between 2010-2012, than in all of our history prior to that

- It is estimated that we are creating **2.5 billion GB per day**
- 30 billion pieces of content shared on Facebook each month
- There are more than 500,000 data centres around the world, large enough to fill about 6,000

How is Big Data generated?

- People spend an increasing amount of time every day consuming information
- Almost all human activities leave a digital trail...
- Electronic transactions on the WWW/Internet
- Continuous **interactions with governments, businesses, organisations (NHS, etc.)**
- Devices, sensors and networks
- **Research (Astronomy, Biology, etc.)**
- **Social media**
- ...
Big data generate on social media & Internet

E-mail

247 billion e-mails a day

Twitter

27 million tweets a day

Blogs: 126 million different blogs


What does Big Data require?

- **Storage**: how do you safely store Big Data
- **Curation**: management and appraisal of data over its lifetime
- **Linkage**: linking (merging) two or more datasets together to produce a single dataset or linking datasets for analysis in real-time

**Data Exploration**

- **Analysis**: using tools to understand and interpret data
- **Visualisation**: providing visual representations of data
- **Modelling**: predictive modelling & hypotheses/theory validation
Using text analysis to make Big Textual Data manageable

- CLASSIFY text so as to identify relevant content / quickly assess this content
- EXTRACT structured information from unstructured data
- SUMMARIZING

TEXT CLASSIFICATION:
SPAM DETECTION

From: *takworld@hotmail.com*
Subject: real estate is the only way... gem - oil vkay

Anyone can buy real estate with no money down
Stop paying rent TODAY !
There is no need to spend hundreds or even thousands for similar courses
I am 22 years old and I have already purchased 6 properties using the methods outlined in this truly INCREDIBLE ebook.
Change your life NOW !
===================================

Click Below to order:
http://www.wholesaledaily.com/sales/nmd.htm
===================================

Dear Hamming Seminar Members

The next Hamming Seminar will take place on Wednesday 25th May and the details are as follows - Who: Dave Robertson
Title: Formal Reasoning Gets Social
Abstract: For much of its history, formal knowledge representation has aimed to describe knowledge independently of the personal and social context in which it is used, with the advantage that we can automate reasoning with such knowledge using mechanisms that also are context independent. This sounds good until you try it on a large scale and find out how sensitive to context much of reasoning actually is. Humans, however, are great hoarders of information and sophisticated tools now make the acquisition of many forms of local knowledge easy. The question is: how to combine this beyond narrow individual use, given that knowledge (and reasoning) will inevitably be contextualized in ways that may be hidden from the people/systems that may interact to use it? This is the social side of knowledge representation and automated reasoning. I will discuss how the formal reasoning community has adapted to this new view of scale. When: 4pm, Wednesday 25 May 2011 Where: Room G07, Informatics Forum There will be wine and nibbles afterwards in the atrium café area.
Id: Abc123 on 5-1-2008 “I bought an iPhone a few days ago. It is such a nice phone. The touch screen is really cool. The voice quality is clear too.

It is much better than my old Blackberry, which was a terrible phone and so difficult to type with its tiny keys. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, ...”
SENTIMENT ANALYSIS

Id: Abc123 on 5-1-2008 “I bought an iPhone a few days ago. It is such a nice phone. The touch screen is really cool. The voice quality is clear too.

It is much better than my old Blackberry, which was a terrible phone and so difficult to type with its tiny keys. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, …”

GALATEAS: EU project on query log analysis

The GALATEAS project offers digital content providers an innovative approach to understanding users' behaviour by analysing language-based information from transaction logs technologies facilitating improved navigation and search for multilingual content access
Case study: The Bridgeman Digital Library

- Collection: 320,000+ images
- Receives over 40,000 queries per month
- Taxonomy: 289 top-categories, 1,148 sub-categories

<table>
<thead>
<tr>
<th>Query</th>
<th>NER TYPE</th>
<th>URI</th>
</tr>
</thead>
<tbody>
<tr>
<td>calling of st. matthew</td>
<td>ARTWORK / OBJ, PERSON</td>
<td><a href="http://en.wikipedia.org/wiki/The_Calling_of_St_Matthew_(Caravaggio)">http://en.wikipedia.org/wiki/The_Calling_of_St_Matthew_(Caravaggio)</a></td>
</tr>
<tr>
<td>joseph</td>
<td>PERSON / SAINT</td>
<td><a href="http://en.wikipedia.org/wiki/St_Joseph">http://en.wikipedia.org/wiki/St_Joseph</a></td>
</tr>
<tr>
<td>george dunlop leslie</td>
<td>PERSON / PAINTER</td>
<td><a href="http://en.wikipedia.org/wiki/George_Dunlop_Leslie">http://en.wikipedia.org/wiki/George_Dunlop_Leslie</a></td>
</tr>
<tr>
<td>the crucible</td>
<td>PLAY / THEATRE</td>
<td><a href="http://en.wikipedia.org/wiki/The_Crucible">http://en.wikipedia.org/wiki/The_Crucible</a></td>
</tr>
<tr>
<td>vesuvius pompeii</td>
<td>LOC, LOC</td>
<td><a href="http://en.wikipedia.org/wiki/Pompeii">http://en.wikipedia.org/wiki/Pompeii</a></td>
</tr>
</tbody>
</table>

Analytics tools

[Analytics tools image]
INFORMATION EXTRACTION: REFERENCES TO (NAMED) ENTITIES

INFORMATION EXTRACTION: FINDING JOBS ON THE WEB

foodscience.com-Job2
JobTitle: Ice Cream Guru
Employer: foodscience.com
JobCategory: Travel/Hospitality
JobFunction: Food Services
JobLocation: Upper Midwest
Contact Phone: 800-488-2611
DateExtracted: January 8, 2001
Source: www.foodscience.com/jobs_midwest.html
OtherCompanyJobs: foodscience.com-Job1
Summarization

• Summarization is the production of a summary either from a single source (single-document summarization) or from a collection of articles (multi-document summarization)

Summarization

– NewsBlaster (Columbia)
  • http://newsblaster.cs.columbia.edu/

A company that acts as a middle man between content companies and Internet service providers is accusing Comcast Corp., the nation’s largest broadband provider, of anti-competitive behavior. (article 1) Comcast Corp. and NBC Universal made new promises to the Federal Communications Commission that the companies hope will help get the regulatory agency to approve the proposed deal between the media giants. (article 2) At issue is the cable operator’s decision to offer the Tennis Channel on a specialty tier of sports networks as opposed to its widely distributed basic tier. (article 3) The quality of television news could deteriorate further under a Comcast-controlled NBC Universal, the Writers Guild of America East warned Wednesday in letters to key Washington officials overruling the government’s review of the proposed merger. (article 4) With regulatory approval still weeks if not months away, Comcast and NBC Universal have extended the terms of their merger agreement to March of next year. (article 5) Patron Michael Copps from the joint venture would put too much control of content into the hands of a company that also controls how consumers access the Internet and television. (article 6) Susan Fox talked on Wednesday with two senior staff members of FCC Commissioner Meredith Averell Baker (article 7)
Contents of the module

- Quick review of preprocessing (today)
- Document clustering and classification
- Sentiment analysis
- Information extraction: Named Entity (NE) recognition & classification, coreference
- Language technology for social media

Outline of lecture

- Text analytics as a solution to the big data problem
- Contents of the module
- **Preprocessing**
  - Existing tools
  - The role of statistics in NLP
  - Module logistics
THE PIPELINE MODEL OF TEXT INTERPRETATION

Quick review of preprocessing

- Language identification
- Tokenization
- Morphological analysis
  - Or at least stemming
- Sentence splitting
- Part of speech tagging
- Parsing
Language identification

Language Identification

• The task of detecting the language a text is written in.

• Identifying the language of a text from some of the text’s attributes is a typical classification problem.

• Two approaches to language identification:
  – Short words (articles, prepositions, etc.)
  – N-grams (sequences of n letters). Best results are obtained for trigrams (3 letters).
Language Identification
Trigram method

• Given a specific language and a text file written in this language, the training module will execute the following steps:
  – Remove characters that may reduce the probability of correct language identification (!, ','(), (){):; ?, & £ § % * 0 1 2 3 4 5 6 7 8 9 . -)
  – Replace all white spaces with _ to mark word boundaries, then replace any sequence of _ with _ so that double spaces are treated as one
  – Store all three-character sequences within an array, with each having a counter indicating number of occurrences
  – Remove from the list of trigrams all trigrams with underscores in the middle (e.g. "_a" for example) as they are considered to be invalid trigrams
  – Retain for further processing only those trigrams appearing more than x times
  – Approximate the probability of each trigram occurring in a particular language by summing the frequencies of all the retained trigrams for that language, and dividing each frequency by the total sum

• This process is repeated for all languages the system should be trained on.
• All language specific trigram data files are merged into one combined training file.
Trigram method
Language detection module

- **Input**: text written in an unknown language
- The unknown text sample is processed in a similar way to the training data (i.e. removing unwanted characters, replacing spaces with underscores and then dividing it into trigrams), and for each trained language the probability of the resulting sequence of trigrams is computed. This assumes that a zero probability is assigned to each unknown trigram.
- The language will be identified by the language trigram data set with the highest combined probability of occurrence.
- The fewer characters in the source text, the less accurate the language detection is likely to be.
- This method is successful in more than 90% of the cases when the input text contains at least 40 characters.

Language Guessers (1)

**SWESUM**
**Author(s)**: Hercules Dalianis, Martin Hassel, KTH, Euroling AB
**Purpose**: Supported languages: Swedish, Spanish, German, French, English
**Access**: Free at [http://www.euroling.se/produkter/swesum.html](http://www.euroling.se/produkter/swesum.html)

**LANGSUITE**
**Author(s)**: PetaMem
**Purpose**: Supported languages: Unicode, Spanish, Polish, Italian, Hungarian, German, French, English, Dutch, Danish, Czech.

**TED DUNNING’S LANGUAGE IDENTIFIER**
**Author(s)**: Ted Dunning
**Access**: Free at [ftp://crl.nmsu.edu/pub/misc/lingdet_suite.tar.gz](ftp://crl.nmsu.edu/pub/misc/lingdet_suite.tar.gz)

**TEXTCAT**
**Author(s)**: Gertjan van Noord
**Purpose**: TextCat is an implementation of the N-Gram-Based Text Categorization algorithm and at the moment, the system knows about 69 natural languages.
**Access**: Free at [http://grid.let.rug.nl/~vannoord/TextCat/](http://grid.let.rug.nl/~vannoord/TextCat/)
Language Guessers (2)

XEROX LANGUAGE IDENTIFIER

Author(s): Xerox Research Centre Europe

Purpose: Supported languages: Albanian, Arabic, Basque, Breton, Bulgarian, Catalan, Chinese, Croatian, Czech, Danish, Dutch, English, Esperanto, Estonian, Finnish, French, Georgian, German, Greek, Hebrew, Hungarian, Icelandic, Indonesian, Irish, Italian, Japanese, Korean, Latin, Latvian, Lithuanian, Malay, Maltese, Norwegian, Polish, Poruguese, Romanian, Russian, Slovakian, Slovenian, Spanish, Swahili, Swedish, Thai, Turkish, Ukrainian, Vietnamese, Welsh


Tokenization
Tokenization

- **Input:** “Friends, Romans and Countrymen”
- **Output:** Tokens
  - *Friends*
  - *Romans*
  - *Countrymen*
- Each such token is now a candidate for an index entry, after further processing
  - Described below
- But what are valid tokens to emit?

Tokenization

- Issues in tokenization:
  - *Finland’s capital*®
    - *Finland? Finlands? Finland’s?*
  - *Hewlett-Packard®*®
    - *Hewlett and Packard*
    - as two tokens?
      - **state-of-the-art**: break up hyphenated sequence.
      - **co-education**
      - **lowercase, lower-case, lower case**?
      - It's effective to get the user to put in possible hyphens
  - *San Francisco*: one token or two? How do you decide it is one token?
Numbers

• 3/12/91  Mar. 12, 1991
• 55 B.C.
• B-52
• My PGP key is 324a3df234cb23e
• (800) 234-2333
  – Often have embedded spaces
  – Often, don’t index as text
    • But often very useful: think about things like looking up error codes/stacktraces on the web
    • (One answer is using n-grams: Lecture 3)
  – Will often index “meta-data” separately
    • Creation date, format, etc.

Tokenization: language issues

• French
  – L'ensemble ☟ one token or two?
    • L ? L' ? Le ?
    • Want l'ensemble to match with un ensemble

• German noun compounds are not segmented
  – Lebensversicherungsgesellschaftsangestellter
  – ‘life insurance company employee’
  – German retrieval systems benefit greatly from a compound splitter module
Tokenization: language issues

• Chinese and Japanese have no spaces between words:
  – 莎拉波娃现在居住在美国东南部的佛罗里达。
  – Not always guaranteed a unique tokenization

• Further complicated in Japanese, with multiple alphabets intermingled
  – Dates/amounts in multiple formats

フォーチュン500社は情報不足のため時間あたり$500K（約6,000万円）

Katakana | Hiragana | Kanji | Romaji
----------|----------|-------|-------

End-user can express query entirely in hiragana!

Tokenization: language issues

• Arabic (or Hebrew) is basically written right to left, but with certain items like numbers written left to right

• Words are separated, but letter forms within a word form complex ligatures

• ‘Algeria achieved its independence in 1962 after 132 years of French occupation.’
• ‘Algeria achieved its independence in 1962 after 132 years of French occupation.’

• With Unicode, the surface presentation is complex, but the stored form is straightforward
Stop words

- With a stop list, you exclude from dictionary entirely the commonest words. Intuition:
  - They have little semantic content: the, a, and, to, be
  - There are a lot of them: ~30% of postings for top 30 wds

- But the trend is away from doing this:
  - Good compression techniques (lecture 5) means the space for including stopwords in a system is very small
  - Good query optimization techniques mean you pay little at query time for including stop words.
  - You need them for:
    - Phrase queries: “King of Denmark”
    - Various song titles, etc.: “Let it be”, “To be or not to be”
    - “Relational” queries: “flights to London”

Normalization

- Need to “normalize” terms in indexed text as well as query terms into the same form
  - We want to match U.S.A. and USA

- We most commonly implicitly define equivalence classes of terms
  - e.g., by deleting periods in a term

- Alternative is to do asymmetric expansion:
  - Enter: window Search: window, windows
  - Enter: windows Search: Windows, windows, window
  - Enter: Windows Search: Windows

- Potentially more powerful, but less efficient
Normalization: other languages

• Accents: résumé vs. resume.
• Most important criterion:
  – How are your users like to write their queries for these words?
• Even in languages that standardly have accents, users often may not type them
• German: Tuebingen vs. Tübingen
  – Should be equivalent

Normalization: other languages

• Need to “normalize” indexed text as well as query terms into the same form
  7月30日 vs. 7/30
• Character-level alphabet detection and conversion
  – Tokenization not separable from this.
  – Sometimes ambiguous:
Case folding

- Reduce all letters to lower case
  - exception: upper case in mid-sentence?
    - e.g., General Motors
    - Fed vs. fed
    - SAIL vs. sail
  - Often best to lower case everything, since users will use lowercase regardless of ‘correct’ capitalization...

- Aug 2005 Google example:
  - C.A.T. → Cat Fanciers website not Caterpillar Inc.

Tokenizers

- Most programming languages provide basic tokenization facilities or regular expressions that can be used for the purpose
  - Java, Perl, Python
- Most machine learning packages provide basic document tokenization facilities to facilitate the development of document classification / clustering
  - SciKit Learn, Vowpal Wabbit, Weka, ...
- All natural language processing platforms (see end of lecture) contain some tokenization facility
  - GATE, LingPipe, NLTK (a Python library), OpenNLP, Stanford CORE
- For large-scale document tokenization, indexing packages such as Elastic Search or Apache SOLR can be used
Morphological analysis / stemming

Lemmatization

- Reduce inflectional/variant forms to base form
- E.g.,
  - am, are, is → be
  - car, cars, car's, cars' → car
- the boy's cars are different colors → the boy car be different color
- Lemmatization implies doing “proper” reduction to dictionary headword form
Stemming

• Reduce terms to their “roots” before indexing
• “Stemming” suggest crude affix chopping
  – language dependent
  – e.g., automate(s), automatic, automation all reduced to automat.

For example compressed and compression are both accepted as equivalent to compress.

Porter’s algorithm

• Commonest algorithm for stemming English
  – Results suggest it’s at least as good as other stemming options
• Conventions + 5 phases of reductions
  – phases applied sequentially
  – each phase consists of a set of commands
  – sample convention: Of the rules in a compound command, select the one that applies to the longest suffix.
Typical rules in Porter

- *sses* $\rightarrow$ *ss*
- *ies* $\rightarrow$ *i*
- *ational* $\rightarrow$ *ate*
- *tional* $\rightarrow$ *tion*

- Weight of word sensitive rules
- $(m>1)$ EMENT $\rightarrow$
  - replacement $\rightarrow$ replac
  - cement $\rightarrow$ cement

Other stemmers

- Other stemmers exist, e.g., Lovins stemmer
  http://www.comp.lancs.ac.uk/compling/research/stonemng/general/lovins.htm
  - Single-pass, longest suffix removal (about 250 rules)
- Full morphological analysis – at most modest benefits for retrieval
- Do stemming and other forms of normalization help?
  - English: very mixed results. Helps recall for some queries but harms precision on others
    - E.g., operative (dentistry) $\Rightarrow$ oper
  - Definitely useful for Spanish, German, Finnish, ...
Off-the-shelf stemming tools

- Implementations of stemming algorithms are included in most NLP packages, including
  - NLTK (see Lab 1)
  - LingPipe, OpenNLP, StanfordCore
- Open source morphological analysis software more rare – Xerox has a professional tool, XFST, that covers most languages but is very expensive

Sentence splitting
Sentence splitting

• Sentence splitting is the task of segmenting text into sentences
• In the majority of cases it is a simple task: . ? ! usually signal a sentence boundary
• However, in cases when a period denotes a decimal point or is a part of an abbreviation, it does not always signal a sentence break.
• The simplest algorithm is known as ‘period-space-capital letter’ (not very good performance). Can be improved with lists of abbreviations, a lexicon of frequent sentence initial words and/or machine learning techniques

Off-the-shelf sentence splitting

• Again, most NLP packages offer some facilities, many can be used in isolation
  – E.g., NLTK, OpenNLP
Television/NN has/HVZ yet/RB to/TO work/VB out/RP a/AT living/RBG arrangement/NN with/IN jazz/NN ./, which/VDT comes/VBZ to/IN the/AT medium/NN more/QL as/CS an/AT uneasy/JJ guest/NN than/CS as/CS a/AT relaxed/VBN member/NN of/IN the/AT family/NN ./.
AMBIGUITY IN POS TAGGING

<table>
<thead>
<tr>
<th>The</th>
<th>AT</th>
</tr>
</thead>
<tbody>
<tr>
<td>man</td>
<td>NN</td>
</tr>
<tr>
<td>still</td>
<td>NN</td>
</tr>
<tr>
<td>saw</td>
<td>NN</td>
</tr>
<tr>
<td>her</td>
<td>PPO</td>
</tr>
<tr>
<td></td>
<td>PP$</td>
</tr>
</tbody>
</table>

POS-Taggers: Frequency + Context

- Most POS-tagger achieve good results by combining
  - FREQUENCY
    - I poured FLOUR/NN into the bowl.
    - Peter should FLOUR/VB the baking tray
  - Information about CONTEXT
    - I saw the new/JJ PLAY/NN in the theater.
    - The boy will/MD PLAY/VBP in the garden.
The importance of context

• Secretariat/NNP is/VBZ expected/VBN to/TO race/VB tomorrow/NN
• People/NNS continue/VBP to/TO inquire/VB the/DT reason/NN for/IN the/DT race/NN for/IN outer/JJ space/NN

Tagging methods

• Hand-coded
• Brill tagger
• Statistical (Markov) taggers
Hand-coded POS tagging: the two-stage architecture

- Early POS taggers all hand-coded
- Most of these (Harris, 1962; Greene and Rubin, 1971) and the best of the recent ones, ENGTWOL (Voutilainen, 1995) based on a two-stage architecture

Hand-coded rules (ENGTWOL)

**STEP 1:** assign to each word a list of potential parts of speech
- in ENGTWOL, this done by a two-lever morphological analyzer (a finite state transducer)

**STEP 2:** use about 1000 hand-coded CONSTRAINTS (if-then rules) to choose a tag using contextual information
- the constraints act as FILTERS
Example

Pavlov had shown that salivation....

Possible tags for `that`

PRON DEM SG: I like THAT
DET CENTRAL DEM SG: I like THAT pen
CS: I said THAT Bill should apologize.
ADV: It isn’t THAT odd.
A constraint

ADVERBIAL-THAT RULE

Given input: “that”

if

(+1 A/ADV/QUANT); /* next word adj. adv. quant */
(+2 SENT-LIM); /* and following that there is a
sentence boundary */
(NOT –1 SVOC/A); /* and previous word is not verb
‘consider’ */

then eliminate non-ADV tags
else eliminate ADV tag.

Markov Model POS tagging

• The problem is to find an ‘explanation’ with
the highest probability:

\[
\arg\max_{t \in \mathcal{T}} P(t_1..t_n | w_1..w_n)
\]

• This can be ‘turned around’ using Bayes’ Rule:

\[
\arg\max_{w_1..w_n} \frac{P(w_1..w_n | t_1..t_n)P(t_1..t_n)}{P(w_1..w_n)}
\]
Combining frequency and contextual information

• This equation can be simplified:

\[
\text{argmax} P(w_1 \ldots w_n \mid t_1 \ldots t_n) P(t_1 \ldots t_n)
\]

• Once further simplifications are applied, this equation will encode both FREQUENCY and CONTEXT INFORMATION

Off-the-shelf POS taggers

• Most packages discussed at the end of this lecture include some POS tagger, often for multiple languages
  – E.g., NLTK contains implementations of / interfaces to several state-of-the-art POS taggers, see http://www.nltk.org/api/nltk.tag.html#module-nltk.tag

• Other POS taggers: see the ACL Wiki
The cat sat on the mat
Where are we now?

Parsing

– Stanford Parser
  (http://nlp.stanford.edu:8080/parser/)

Off-the-shelf parsers

• Many included in off-the-shelf packages
  – NLTK
  – OpenNLP (OpenCCG)

• Others listed at
Outline of lecture

- Text analytics as a solution to the big data problem
- Contents of the module
- Preprocessing
- **Existing tools**
  - The role of statistics in NLP
  - Module logistics

Off-the-shelf NLP packages
Off-the-shelf NLP packages

- NLTK
- Stanford NLP
- GATE
- SOLR
- UIMA
- Others

The Natural Language Toolkit

- [http://www.nltk.org/](http://www.nltk.org/)
- A suite of Python modules for natural language processing
- Includes modules for:
  - Classification
  - Parsing
  - Tokenization
  - Stemming
  - Tagging
  - NE recognition and other Information Extraction
Stanford NLP Tools

- The Stanford NLP Research group offers a series of open-source NLP tools for text manipulation, implemented in Java:
  - The Stanford Parser: probabilistic natural language parsers
  - The Stanford POS Tagger: a maximum-entropy POS tagger
  - The Stanford Named Entity Recognizer: features for Named Entity Recognition
  - The Stanford Classifier: conditional loglinear classifier
  - Topic Modeling Toolbox: a suite of topic modeling tools

GATE

- [http://gate.ac.uk](http://gate.ac.uk)
- A comprehensive open source infrastructure for developing language processing applications
- Written in Java
- Mature and actively supported
GATE components

- Provides a baseline set of customizable Language Engineering components that can be extended and/or replaced by users, for the following NLP tasks:
  - Tokenization
  - POS tagging
  - Sentence splitting
  - Named entity recognition
  - Co-reference resolution
  - Information Extraction
  - Machine learning, etc

Apache SOLR

- SOLR:
  - [http://lucene.apache.org/solr](http://lucene.apache.org/solr)
  - An indexing and searching library
  - Provides the ability to preprocess documents in many different languages
Unstructured Information Management Architecture (UIMA)

- [http://uima.apache.org](http://uima.apache.org)
- Enables the creation and aggregation of single NLP tools (called Analysis Engines (AEs)) into pipelines (aggregate AEs)
- Developed by IBM, now part of Apache
- Available for Java and C++, but supports also components in Perl, Python, and TCL

UIMA Components

- Current annotators available for UIMA include:
  - Tokenizers
  - Sentence Splitter
  - Stemmers
  - Acronym Annotator
  - Named Entity Tagger
  - Lucene Indexer
  - Concept Mapper
  - Feature Extractor, etc.
Others

• OpenNLP: http://opennlp.apache.org/
• Lingpipe: http://alias-i.com/lingpipe/

Outline of lecture

• Text analytics as a solution to the big data problem
• Contents of the module
• Preprocessing
• Existing tools
• The role of statistics in NLP
• Module logistics
Statistical methods in text analysis

Using machine learning for text analysis

- Most language interpretation task require the ability to handle an almost infinite number of variations and to decide between a high number of distinct hypotheses
- It is very difficult for a programmer to be able to think of all the possibilities and come up with good decision criteria
- For this reason although some very high-performing hand-coded systems exist, most text analysis tools use some form of machine learning to learn how to do the task
WSD: SENSES OF “line”

- **Product:** “While he wouldn’t estimate the sale price, analysts have estimated that it would exceed $1 billion. Kraft also told analysts it plans to develop and test a line of refrigerated entrees and desserts, under the Chilery brand name.”

- **Formation:** “C-LD-R L-V-S V-NNA reads a sign in Caldor’s book department. The 1,000 or so people fighting for a place in line have no trouble filling in the blanks.”

- **Text:** “Newspaper editor Francis P. Church became famous for a 1897 editorial, addressed to a child, that included the line “Yes, Virginia, there is a Santa Clause.”

- **Cord:** “It is known as an aggressive, tenacious litigator. Richard D. Parsons, a partner at Patterson, Belknap, Webb and Tyler, likes the experience of opposing Sullivan & Cromwell to ‘having a thousand-pound tuna on the line.’”

- **Division:** “Today, it is more vital than ever. In 1983, the act was entrenched in a new constitution, which established a tricameral parliament along racial lines, with separate chambers for whites, coloreds and Asians but none for blacks.”

- **Phone:** “On the tape recording of Mrs. Guba’s call to the 911 emergency line, played at the trial, the baby sitter is heard begging for an ambulance.”

---

A SPATIAL VIEW OF CLASSIFICATION

![SPAM](image1)

![NON-SPAM](image2)
A SPATIAL VIEW OF LEARNING

The task of the learner is to learn a function that divides the space of examples into black and red.

LEARNING A FUNCTION

- Given a set of input/output pairs, find a function that does a good job of expressing the relationship:
  - Wordsense disambiguation as a function from words (the input) to their senses (the outputs)
  - Categorizing email messages as a function from emails to their category (spam, useful)
WAYS OF LEARNING A FUNCTION

• SUPERVISED: given a set of example input / output pairs, find a rule that does a good job of predicting the output associated with an input
• UNSUPERVISED learning or CLUSTERING: given a set of examples, but no labelling, group the examples into “natural” clusters

SUPERVISED LEARNING OF WORD SENSE DISAMBIGUATION

• Training experience: a CORPUS of examples annotated with their correct wordsense (according, e.g., to WordNet)
UNSUPERVISED LEARNING OF WORDSENSE DISAMBIGUATION

• Identify distinctions between contexts in which a word like LINE is encountered and hypothesize sense distinctions on their basis
  – Cord: fish, fishermen, tuna
  – Product: companies, products, finance
• (The way children learn a lot of what they know)

Outline of lecture

• Text analytics as a solution to the big data problem
• Contents of the module
• Preprocessing
• Existing tools
• The role of statistics in NLP
• Module logistics
Module logistics

• The teachers:
  – Massimo Poesio, poesio@essex.ac.uk
  – GTA: Dimitrios Andreou, proj.andreou@gmail.com

• Web page with module materials and timetable:
  – http://csee.essex.ac.uk/staff/poesio/Teach/807/

• Assessment:
  – 2 assignments (1 text classification/sentiment analysis, 1 information extraction)
  – Exam

Readings

• Practical intros to the topics of this module
  – Using Python:
    • Richert and Coelho - Building Machine Learning Systems with Python (2nd ed) - Pack Press (RC)
  – Using Hadoop / SOLR / Mahout:
    • G. S. Ingersoll, T. S. Morton, & A. L. Farris – Bringing text – Manning, 2013

• A more theoretical intro to the main topics of the module:

• A general intro to NLP with Python with a practical bend:
  – S. Bird, E. Klein & E. Loper, Natural Language Processing with Python, O’Reilly

• A more theoretically-oriented general reference on NLP:

• A very practical intro to preprocessing and to SOLR:
  – M. Morris – Text Processing in Java – Colloquial Media, 2014

• There is hundreds of textbooks on Text Mining, one I like is