SNS COLLEGE OF ENGINEERING

Kurumbapalayam (Po), Coimbatore – 641 107

An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A' Grade

Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

COURSE NAME : 19CS732 INFORMATION RETRIEVAL TECHNIQUES

IVYEAR / VIII SEMESTER

Unit 3- TEXT CLASSIFICATION AND CLUSTERING

Topic 1 : A Characterization of Text Classification and Unsupervised Algorithms: Clustering , Naïve Text Classification

A Characterization of Text Classification and Unsupervised Algorithms: Clustering , Naïve Text Classification - **Problem**

- Planning matters, but so does flexibility
- Try maintaining an overview of the modeling workflow

- Given an example, classify if it is spam or not.
- Given a handwritten character, classify it as one of the known characters.
- Given recent user behavior, classify as churn or not.

What is Text-Mining?

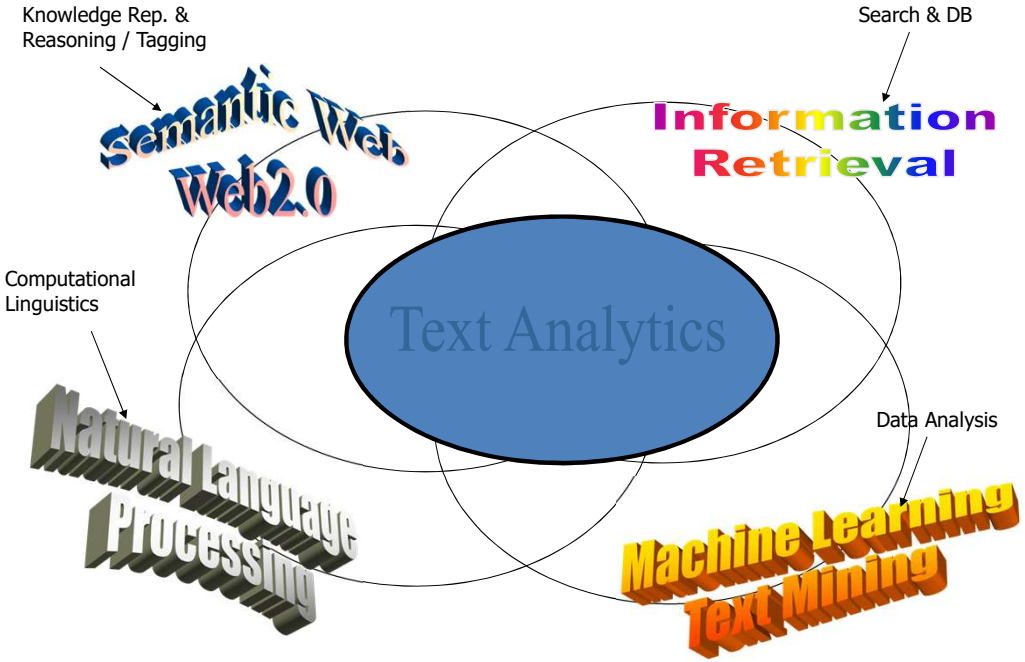
What is Text-Mining?

- finding **interesting** regularities in large **textual** datasets…" (adapted from Usama Fayad)
  - …where **interesting** means: non-trivial, hidden, previously unknown and potentially useful
- "…finding semantic and abstract information from the surface form of textual data…"

Text-Mining-Cont..

- Abstract concepts are **difficult to represent**
- **"Countless" combinations** of subtle, abstract relationships among concepts
- **Many ways** to represent similar concepts
  - E.g. space ship, flying saucer, UFO
- Concepts are **difficult to visualize**
- **High dimensionality**
- **Tens or hundreds of thousands of features**

Who is in the text analysis arena?

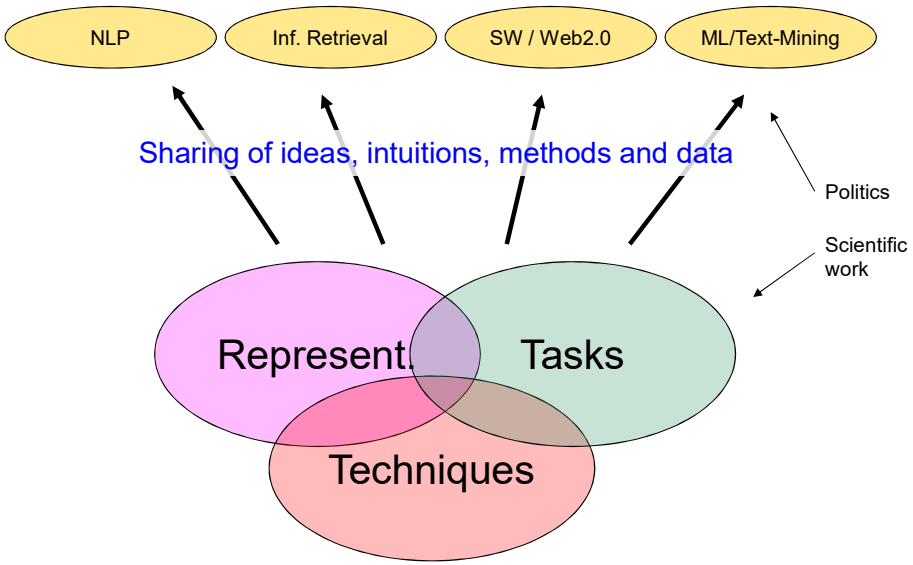


What dimensions are in text analytics?

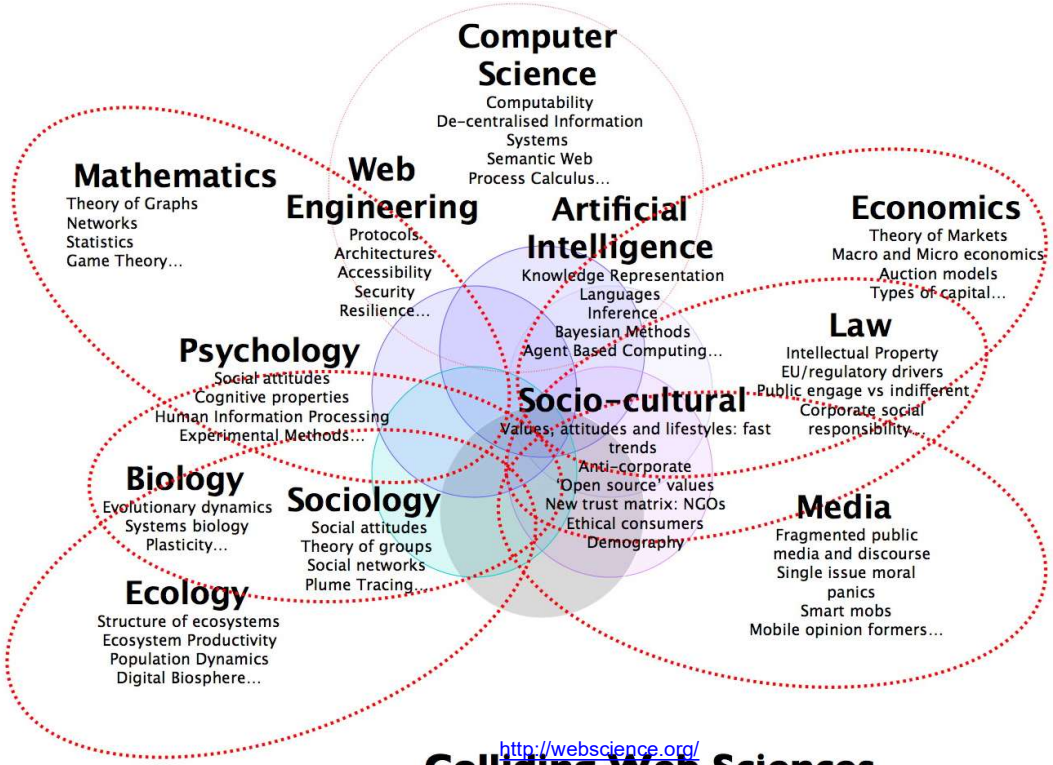
Three major dimensions of text analytics:

- Representations
  - …from character-level to first-order theories
- Techniques
  - …from manual work, over learning to reasoning
- Tasks
  - …from search, over (un-, semi-) supervised learning, to visualization, summarization, translation …

How dimensions fit to research areas?



Broader context: Web Science



Text-Mining
How do we represent text?

Character (character n-grams and sequences)

Words (stop-words, stemming, lemmatization)

Phrases (word n-grams, proximity features)

Part-of-speech tags

Taxonomies / thesauri

Vector-space model

Language models

Full-parsing

Cross-modality

Collaborative tagging / Web2.0

Templates / Frames

Ontologies / First order theories

Character level

- Character level representation of a text consists from sequences of characters…
  - …a document is represented by a frequency distribution of sequences
  - Usually we deal with contiguous strings…
  - …each character sequence of length 1, 2, 3, … represent a feature with its frequency

Word Level

- The most common representation of text used for many techniques
  - …there are many tokenization software packages which split text into the words
- Important to know:
  - Word is well defined unit in western languages – e.g. Chinese has different notion of semantic unit

Word Level –Cont..

- Relations among word surface forms and their senses:
  - **Homonomy**: same form, but different meaning (e.g. bank: river bank, financial institution)
  - **Polysemy**: same form, related meaning (e.g. bank: blood bank, financial institution)
  - **Synonymy**: different form, same meaning (e.g. singer, vocalist)
  - **Hyponymy**: one word denotes a subclass of an another (e.g. breakfast, meal)
- Word frequencies in texts have **power distribution**:
  - …small number of very frequent words
  - …big number of low frequency words.

Part-of-Speech level

- By introducing part-of-speech tags we introduce word-types enabling to differentiate words functions
  - For text-analysis part-of-speech information is used mainly for “information extraction” where we are interested in e.g. named entities which are “noun phrases”
  - Another possible use is reduction of the vocabulary (features)
    - …it is known that nouns carry most of the information in text documents
- Part-of-Speech taggers are usually learned by HMM algorithm on manually tagged data

Part-of-Speech level –Cont..

part of speech	function or "job"	example words	example sentences
Verb	action or state	(to) be, have, do, like, work, sing, can, must	EnglishClub.com **is** a web site. I **like** EnglishClub.com.
Noun	thing or person	pen, dog, work, music, town, London, teacher, John	This is my **dog**. He lives in my **house**. We live in **London**.
Adjective	describes a noun	a/an, the, 69, some, good, big, red, well, interesting	My dog is **big**. I like **big** dogs.
Adverb	describes a verb, adjective or adverb	quickly, silently, well, badly, very, really	My dog eats **quickly**. When he is **very** hungry, he eats **really** quickly.
Pronoun	replaces a noun	I, you, he, she, some	Tara is Indian. **She** is beautiful.
Preposition	links a noun to another word	to, at, after, on, but	We went **to** school **on** Monday.
Conjunction	joins clauses or sentences or words	and, but, when	I like dogs **and** I like cats. I like cats **and** dogs. I like dogs **but** I don't like cats.
Interjection	short exclamation, sometimes inserted into a sentence	oh!, ouch!, hi!, well	**Ouch**! That hurts! **Hi**! How are you? **Well**, I don't know.

Activity

Disadvantages

- **Slow:** For larger dataset, it requires a large amount of time to process.
- **Poor performance with Overlapped classes** : Does not perform well in case of overlapped classes.
- **Selecting appropriate hyperparameters is important:** That will allow for sufficient generalization performance.
- **Selecting the appropriate kernel**

Advantages

- Performs well in Higher dimension
- Best algorithm when classes are separable
- Outliers have less impact.
- SVM is suited for extreme case binary classification.

Assessment 1

1. List out the Advantages of text Classification
   a)_____
   b)_____
   c)_____
   d)_____

2. Identify the disadvantages of text Classification
   a)_____
   b)_____
   c)_____
   d)_____

TEXT BOOKS:

1. Ricardo Baeza-Yates and Berthier Ribeiro-Neto, —Modern Information Retrieval: The Concepts and Technology behind Search, Second Edition, ACM Press Books, 2011.

2. Ricci, F, Rokach, L. Shapira, B.Kantor, —Recommender Systems Handbook||, First Edition, 2011.

REFERENCES:

1. C. Manning, P. Raghavan, and H. Schütze, —Introduction to Information Retrieval, Cambridge University Press, 2008.

2. Stefan Buettcher, Charles L. A. Clarke and Gordon V. Cormack, —Information Retrieval: Implementing and Evaluating Search Engines, The MIT Press, 2010.

THANK YOU