APAN PS5430
Applied Text & Natural Language Analytics
Week 1

Javid Huseynov, Ph.D.
Thursday, January 23, 2020

COLUMBIA UNIVERSITY

Week 1 Agenda

- Introductions
  - Course Instructor & Associate
  - Overview of Natural Language Processing
  - Why NLP & Text Analytics?
- Learning Objectives
- Development Tools
- Reading Materials
- Deliverables
- Python Language

Brief Intro

Instructor

Javid Huseynov, Ph.D.

Principal Data Scientist, STSM
IBM Corporation

- Natural Language Processing
- Distributed Computing
- Machine Learning
- Signal Processing
- Web Design Visualization
- Embedded Systems

Associate

John Ortega

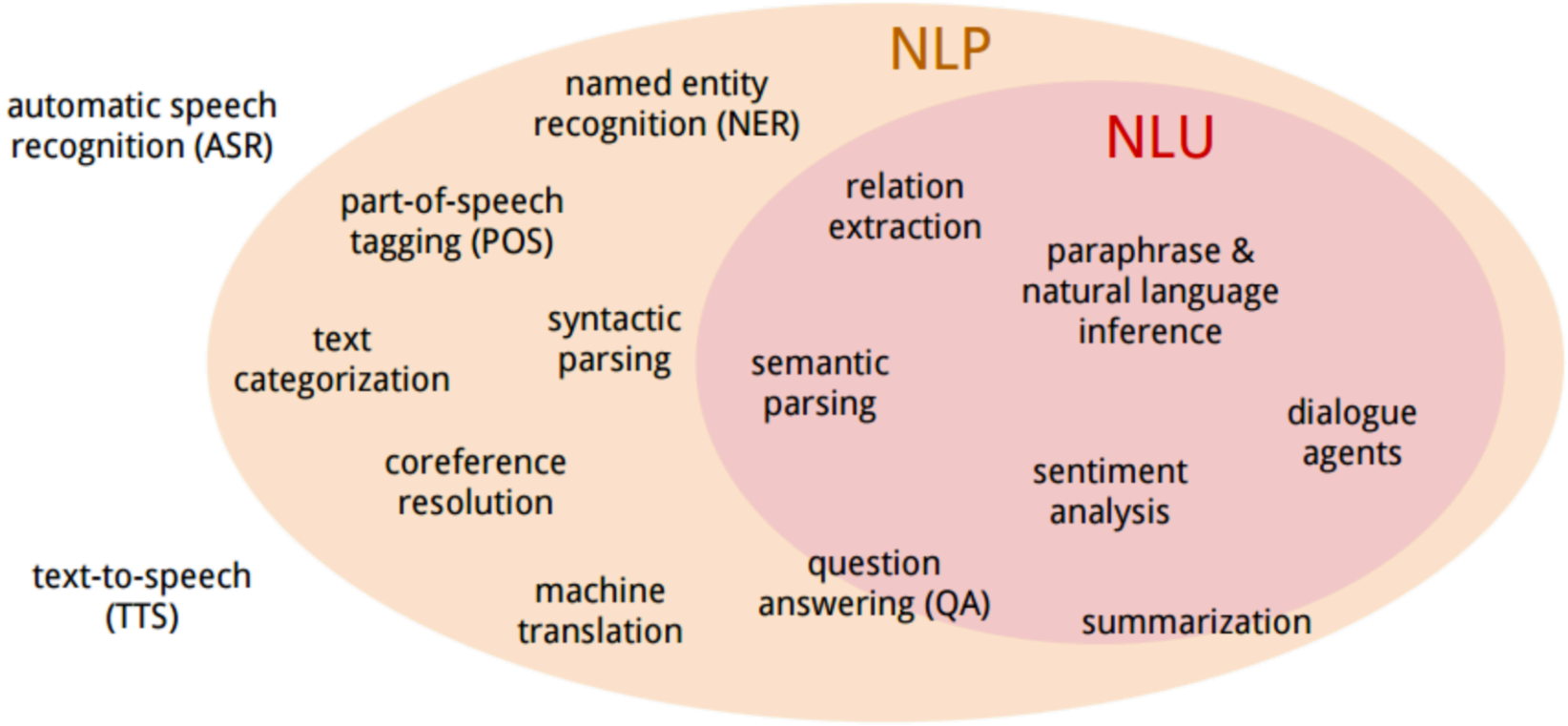
Ph.D. Candidate, Machine Translation
University of Alicante, Spain

- Natural Language Processing
- Machine Translation
- Machine Learning
- Software Engineering
- Web Development
- Management

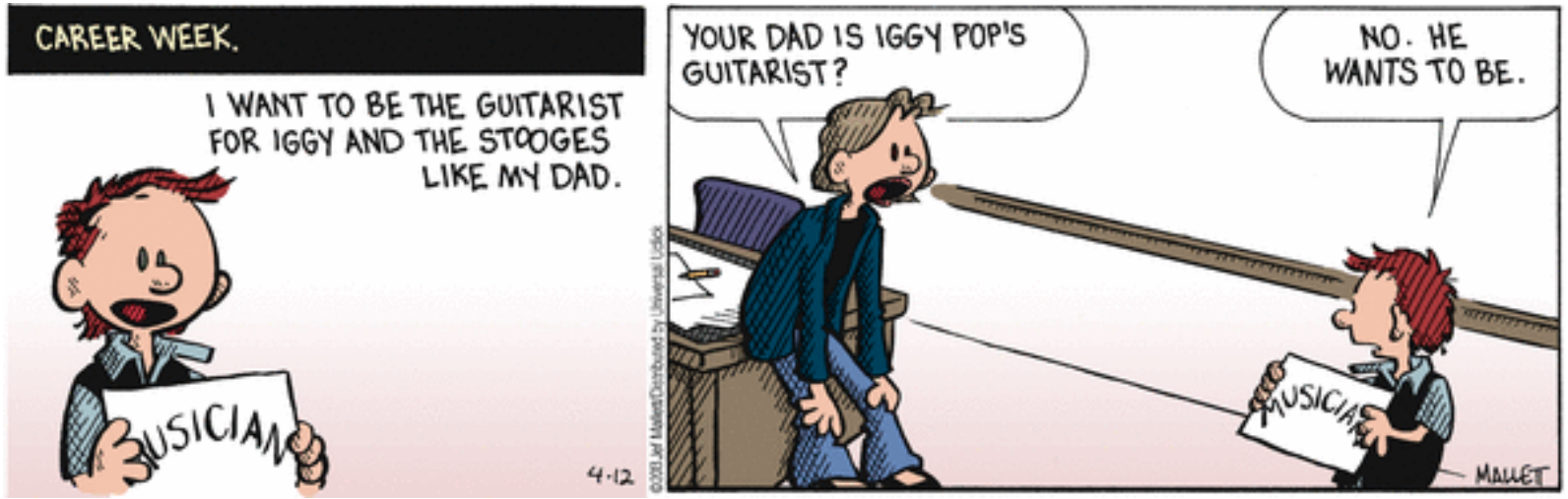
Introduction: IBM Debater Teaser

NLP: Definition & Overview

Natural language processing (**NLP**) is a subfield of computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data.



The ambiguity of language



Non-standard meaning & complexity of languages

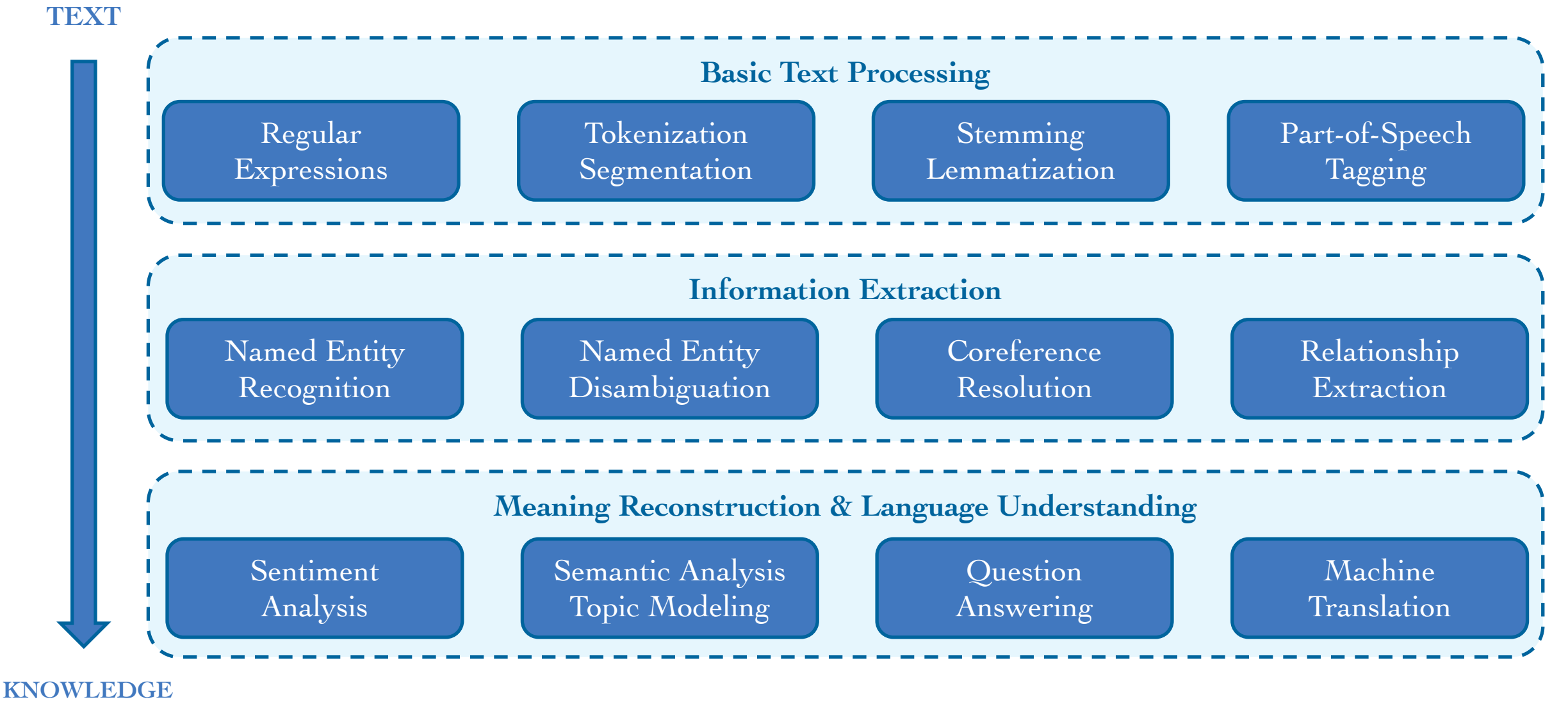
- There are around **7,000 languages and dialects** in the world.
- German: Donaudampfschiffahrtsgesellschaftskapitän (5 "words")
- Chinese: 50,000 different characters (2-3k to read a newspaper)
- Japanese: 3 writing systems
- Thai: ambiguous word boundaries and sentence concepts
- Slavic: Different word forms depending on gender, case, tense

"There it is, 'Twerk', right next to 'Twerp'."

© Brian Crane.

- Neologisms
- Complex entity names
- Phrasal verbs/idioms

NLP Pipeline Tasks



Applied Text Analytics: Example Use Case

- Discover insights about business entities (companies) from publicly-available unstructured data

Firmographics

Unstructured Data

NLP Challenges

ENTITY RESOLUTION

EVOLING TOPIC EXTRACTION

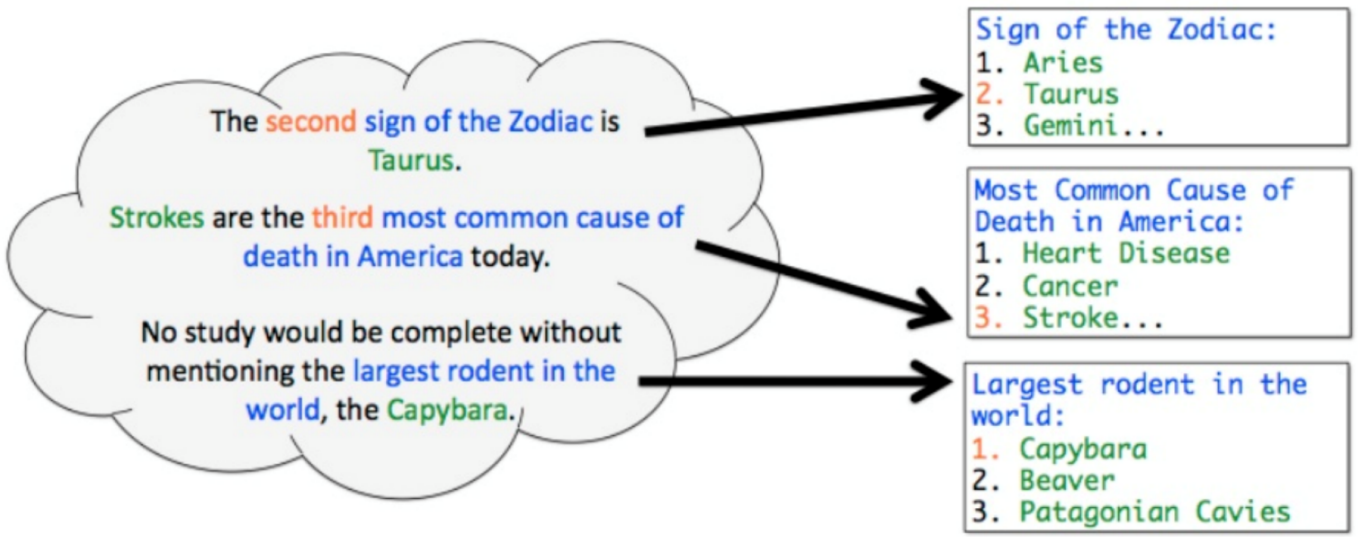
Use Cases	Benefit
Sales & Marketing	Relevant product recommendations for companies
Market Analysis	Identify trends for company or market segment growth
Trust & Compliance	Aggregate adverse information about companies
Client Credit Financing	Discover insights for the company risk assessment

- **Challenges - Develop AI-based approaches to:**
  1. Resolve company mentions in text against a company knowledge base record, **or**
  2. Extract evolving topics & sentiment for selected companies over a time range

Information Extraction (IE)

The NLP task of **extracting structured** (semantic) **information** from unstructured text, to enable:

- further modeling by computer algorithms
- meaning and knowledge extraction
- language understanding



Subtasks

- Named Entity Recognition
- Named Entity Linking
- Coreference Resolution
- Relationship Extraction

Tools

- IBM Watson NLU
- Google Cloud NL
- Amazon Comprehend
- Thomson Reuters Open Calais
- Microsoft Text Analytics
- Stanford CoreNLP*
- spaCy*
- Natural Language Toolkit (NLTK)*

*** open-source**

Challenge: Entity Resolution, Linking or Disambiguation

Shares in Pandora Soar

2018-10-12
COPENHAGEN, Denmark âÂ Shares in Pandora A/S have soared 12 percent in less than two days, putting hedge funds betting against the Danish jewellery maker in a potentially awkward position.
For the past year, Pandora has tended to deliver gains to funds shorting it and losses to long-term investors. In the 12 months through Oct. 10, the stock plunged more than 40 percent amid concerns the company was struggling in the US and China. But this week, something changed.
On Thursday, Pandora suddenly shot up, rising more than 10 percent at one point. On Friday, the shares traded as much as 7.4 percent higher. The gains follow speculation that the company may have hired an adviser to help it handle potential approaches from buyout funds. It also appears to be doing much better in the US, according to a Carnegie note seen by Bloomberg.

Pandora: Danish Jewelry Maker

Pandora names Brad Minor as head of brand marketing and communications

2018-10-26
Streaming audio company Pandora has appointed Brad Minor as vice president, head of brand marketing and communications. He will be responsible for defining Pandora âs brand positioning to listeners and the industry together with the brand marketing, communications, experiential and social media teams.
Minor joins Pandora after serving as managing director of brand, advertising and media at JP Morgan Chase, where he and his team launched the Chase Sapphire Card travel campaigns starring James Corden. Prior to Chase, Minor had roles leading brand marketing, public relations, experiential marketing, social media and partnerships at American Express, focusing on the companyâs music, sports, small business and digital platforms.
âOver the past decade, working with some of the most successful and well-

Pandora: U.S. Music Media Company

Disambiguating business entity mentions in text against company knowledge base(s), e.g. Crunchbase, Wikidata, OpenCorporates, etc.

Semantic Analysis & Information Retrieval (IR)

- **Semantic Analysis** is the task of building structures that *approximate* **meaning** from text
- Related areas
  - Word Sense Disambiguation
  - Keyword Extraction
  - Topic Modeling
  - Question Answering
  - Sentiment Analysis
- **Information Retrieval a.k.a. Search** is the task of obtaining information relevant to an input query

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

Semantic Analysis Approaches

- Vector-space models based on word co-occurrence in corpora
- Dependency models based on syntactic relations in text

Methods

- Lexicon-based
- Statistical or ML-based

IR Models

- Set-theoretical
  - Boolean, Fuzzy
- Algebraic
  - Vector-Space Model, LSA
- Probabilistic
  - Bayesian, LDA

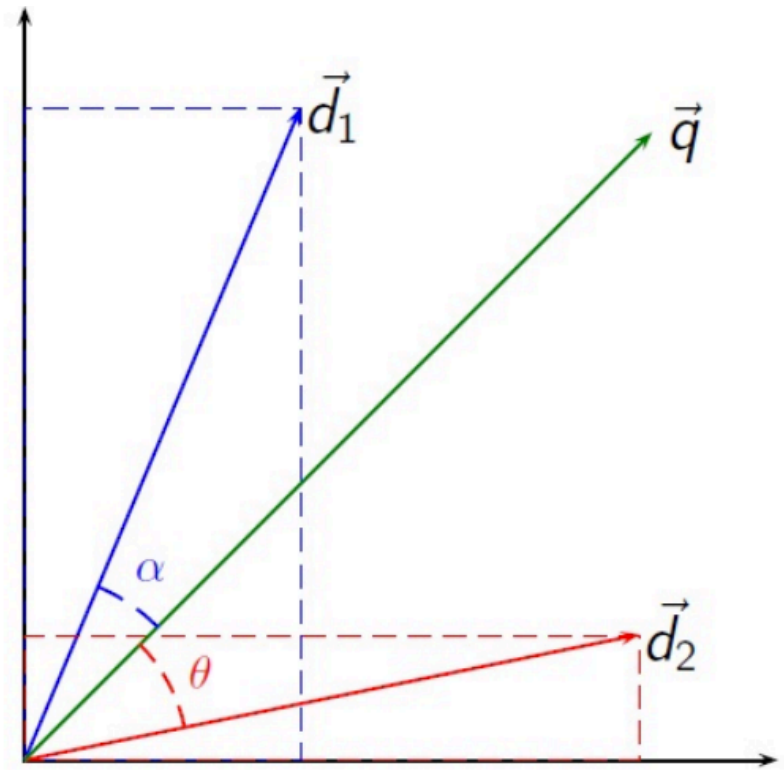
Vectors Space Model

- An algebraic representation of text documents or queries as vectors

$$d_j = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$$
$$q = (w_{1,q}, w_{2,q}, \ldots, w_{n,q})$$

- Similarity measure – Cosine of the angle between vectors

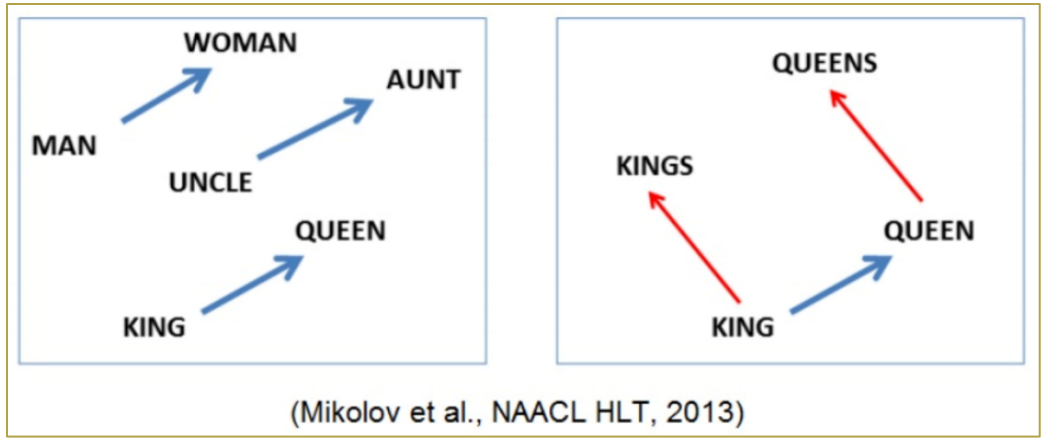
$$\mathbf{sim}(d_j, q) = \frac{\mathbf{d_j} \cdot \mathbf{q}}{\|\mathbf{d_j}\| \, \|\mathbf{q}\|} = \frac{\sum_{i=1}^{N} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2} \sqrt{\sum_{i=1}^{N} w_{i,q}^2}}$$

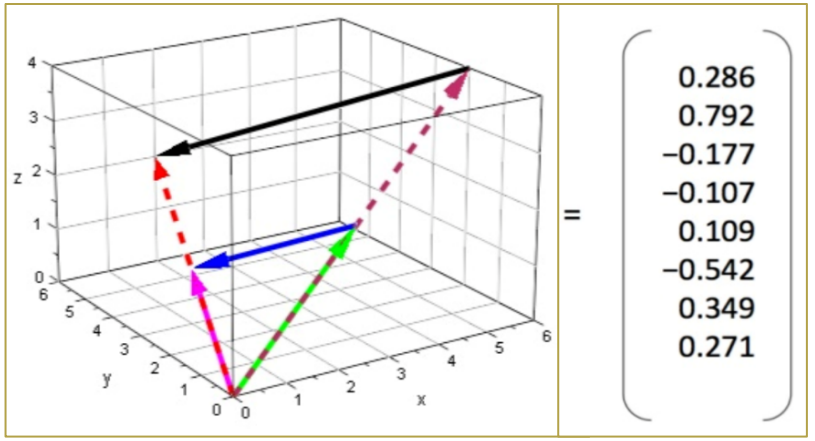


$$\cos\theta = \frac{\mathbf{d_2} \cdot \mathbf{q}}{\|\mathbf{d_2}\| \, \|\mathbf{q}\|} \qquad \|\mathbf{q}\| = \sqrt{\sum_{i=1}^{n} q_i^2}$$

Word Vectors

- Vectors are directions in space, which can also encode relationships:
  - e.g. **man** is to **woman** as **king** is to **queen**





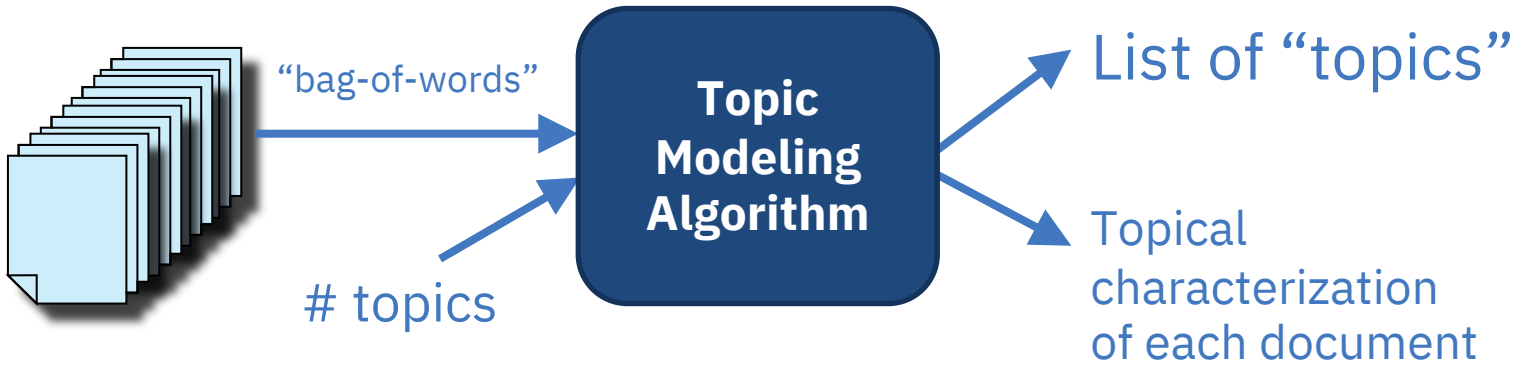
- One of the most successful ideas in modern statistical NLP

  - …government debt problem turning into banking crises as has happened in…
  - …saying that Europe needs unified banking regulation to replace the hodgepodge…

surrounding words capture the context of the word *banking*

Topic Modeling

- Unsupervised method to organize, understand, summarize a set of documents
  - Discovering hidden topical patterns that are present across the collection
  - Annotating documents according to these topics
  - Using these annotations to organize, search and summarize texts

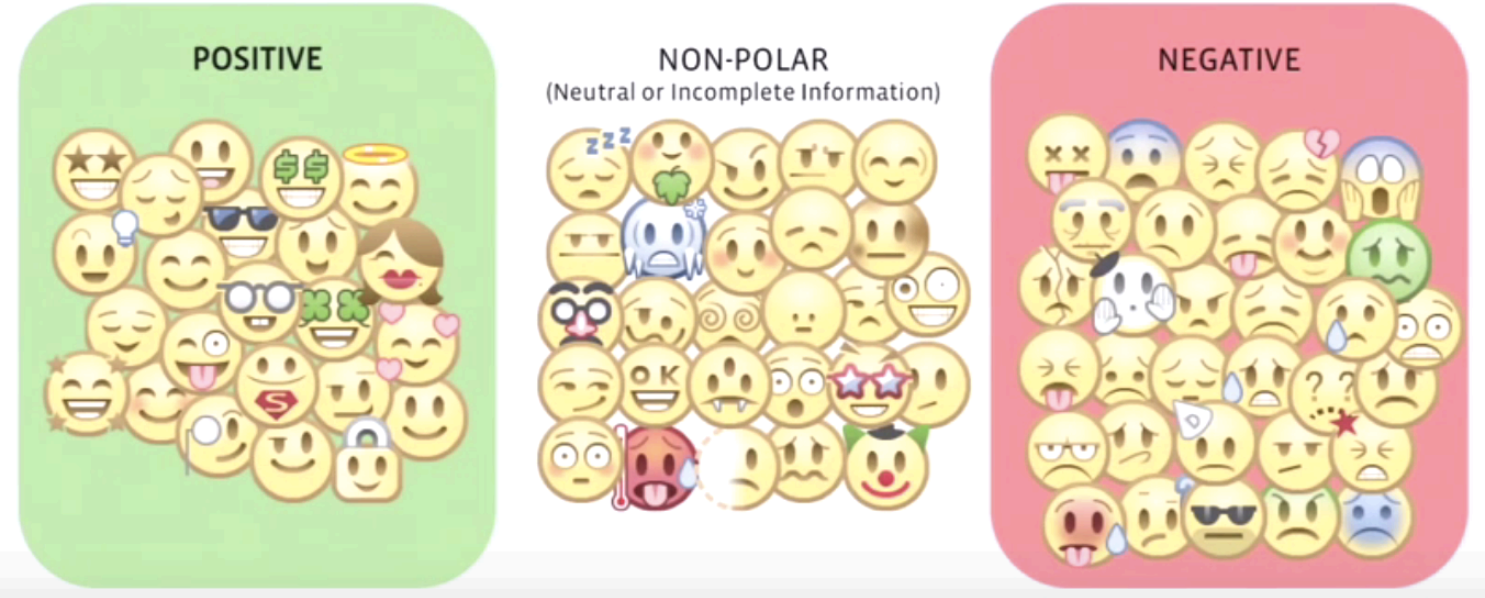


- Useful applications:
  - Improved web searching
  - Auto-indexing of archives
  - Specialized search browsers
  - Legal applications

Sentiment Analysis, a.k.a. Opinion Mining

The task of computationally *identifying* and *categorizing* **opinion** expressed in text, according to the following classes:

- **polarity**: positive, negative, neutral
- **subjectivity**: subjective vs objective



Methods

- Knowledge- or Lexicon-based
- Statistical or Machine Learning
- Hybrid Lexical & ML

Applications

- Social Media
- Product Reviews & Feedback
- Opinion Polls
- Recommender Systems

Tools

- IBM Watson NLU
- Google Cloud NL
- Amazon Comprehend
- NLTK Sentiment Analyzer

Course Learning Objectives

- L1: Derive structured data representation from an unstructured text
- L2: Extract entities and their contexts (relations, keywords, concepts, etc.) from unstructured text
- L3: Model meaning or natural language understanding from text corpora
- L4: Develop descriptive, predictive, and prescriptive analytics models based on text
- L5: Develop opinion / sentiment analysis model based on text corpora

Development Tools

required

optional

Reading Materials

- Bengfort, B. (2018). *Applied text analysis with Python: Enabling language-aware data products with machine learning*. O'Reilly Media, Inc. ISBN: 1491963042 (Hereafter referred to as "ATAP")

- Hapke, H., Howard, C., & Lane, H. (2019). *Natural language processing in action* (Links to an external site.). Manning Publications. ISBN: 9781617294631 (Hereafter referred to as "NLPA")

Available online:

- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media, Inc. ISBN: 9780596550967. Also published as *Analyzing text with the natural language toolkit (Links to an external site.)*. (Hereafter referred to as "NLTK")

- Joshi, P. (2018). An introduction to text summarization using the TextRank algorithm (with Python implementation) (Links to an external site.). Analytics Vidhya. Nov. 1, 2018.

- Jurafsky, D., & Martin, J. (2018). *Speech and language processing (Links to an external site.)*. Third Edition Draft. (Hereafter referred to as "SLP")

Course Deliverables

- Attendance & Participation: 5%
- 9 Weekly Assignments: 45%
- Group Term Project: 40%
  - Term Project Deliverable 1: Term Project Proposal
  - Term Project Deliverable 2: Python Implementation
  - Term Project Deliverable 3: Final Presentation
- Final Reflection + Project Summary: 10%

Python Language Session