Introduction to Text Analytics

Kevin Lee

Department of Statistics
Western Michigan University

Natural Language Processing (NLP)

- Wikipedia says,

  Natural language processing (NLP) is a subfield of linguistics, 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.

- We will focus on Text Analytics!

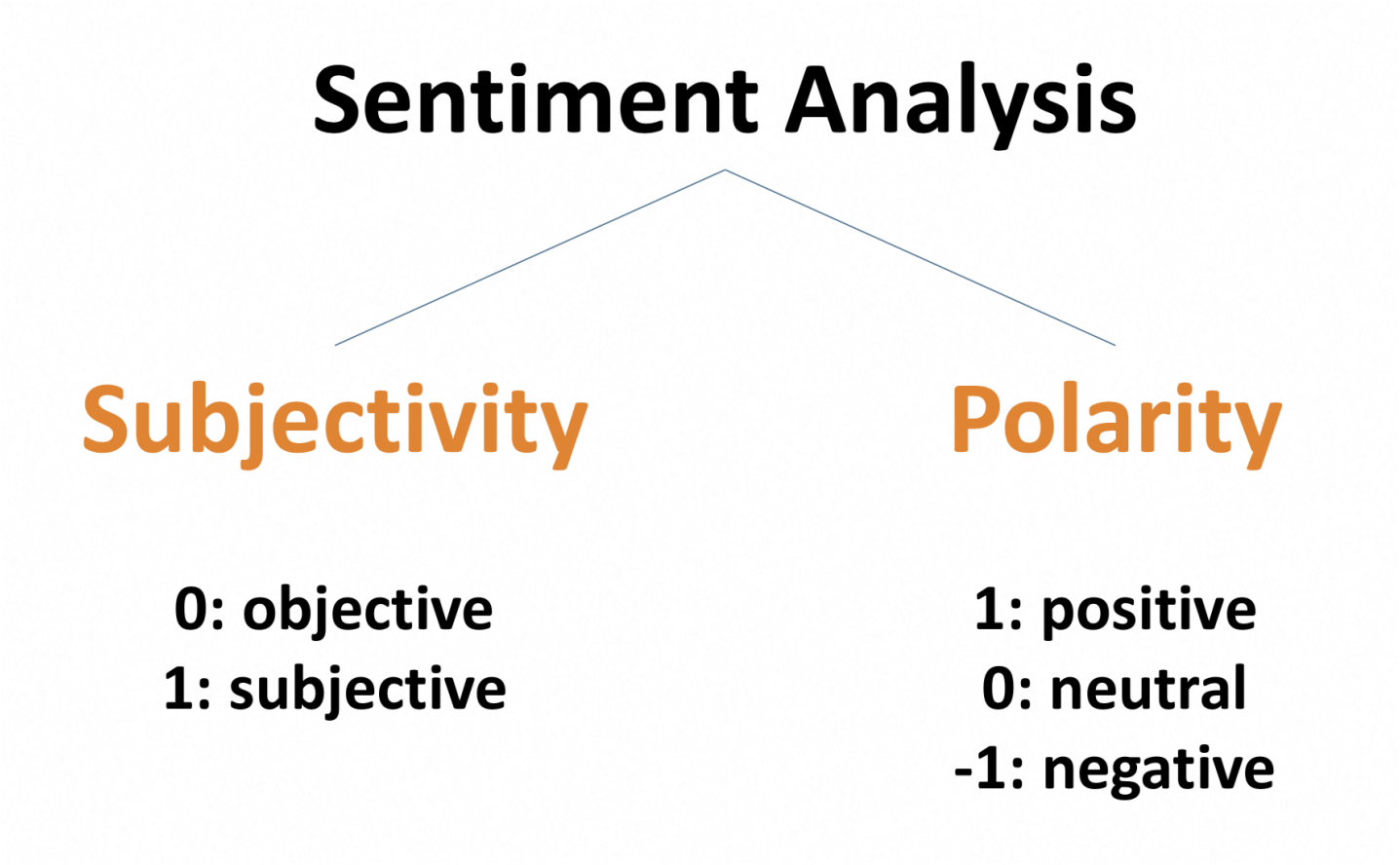
Text Analytics

- Text analytics is used for finding structures or extracting information from unstructured text.

- Text analytics is challenging due to the nature of language.
  - "running is good for your health"
  - "running for office is difficult"

Text Analytics

Parsing level	Vs.	Application level
Tokenization		Text summarization
Lemmatization		Text classification
Part-of-speech (PoS) tagging		Text clustering
Dependency parsing		Topic modeling
		Sentiment analysis (Opinion mining)
		Language detection
		Machine translation

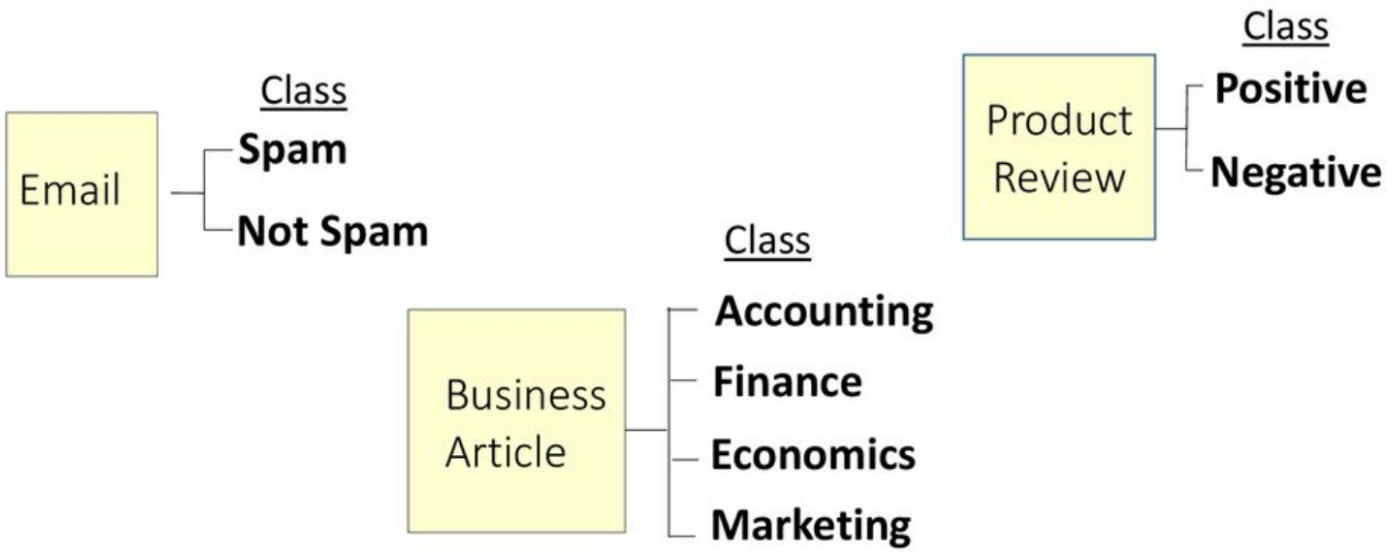

Sentiment Analysis
Subjectivity
0: objective
1: subjective
Polarity
1: positive
0: neutral
-1: negative

Text Analytics - Sentiment Analysis

- Sentiment analysis seems easy, but actually is challenging, often even to humans.
  - Negation & Double Negation
  - Multiple sentiments in a sentence
  - Sarcasm
  - Ambiguity

Text Analytics - Text Classification

- Text classification is the process of assigning a labeled category, known as a class, to text.



Text Analytics - Topic Modeling

- Topic modeling is one of the methods for text clustering.

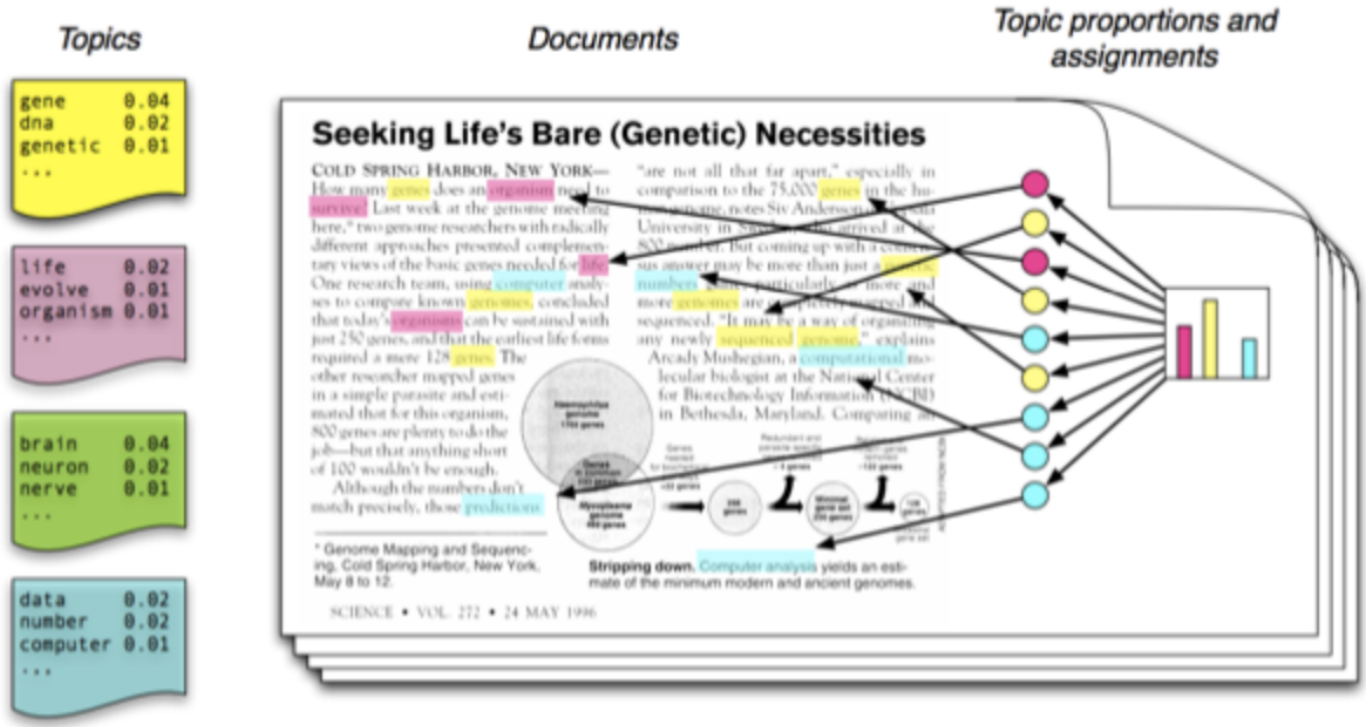


Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

Potential Topics for Final Project

- Sentiment Analysis / Text summarization
  - `NLTK` package
  - `spaCy` package
- Text Classification (Feature Engineering + Machine Learning Methods)
  - Count vectors as features
  - TF-IDF vectors as features
- Topic Modeling
  - LDA (Latent Dirichlet Allocation)
  - NMF (Non-negative Matrix Factorization)
  - `Gensim` package