Advanced Text Analysis for Business (IDS-566)

Lecture 1

Jan 19, 2018

Course Overview

- Instructor
  - Ehsan M. Ardehaly PhD, ehsan@uic.edu
  - Office hours: TBA
  - Teacher assistant: Ramah Al Balawi, ralbal2@uic.edu
- Objectives:
  - Mining patterns from text
  - Study of machine learning concepts
  - Design and implementation of text mining approaches
  - Applications for business decisions

Course Overview

- Suggested text books:
  - Fundamentals of Predictive Text Mining (2nd Edition), Sholom M. Weiss, Nitin Indurkhya, Tong Zhang, 2015
  - Mining Text Data, Charu C. Aggarwal and ChengXiang Zhai, Springer, 2012
  - Mining of Massive Datasets, Jure Leskovec, Anand Rajaraman, Jeff Ullman

- Grading:
  - Final exam: 40%
  - 3 Assignments: 60% (3 x 20%)

Course Assignments

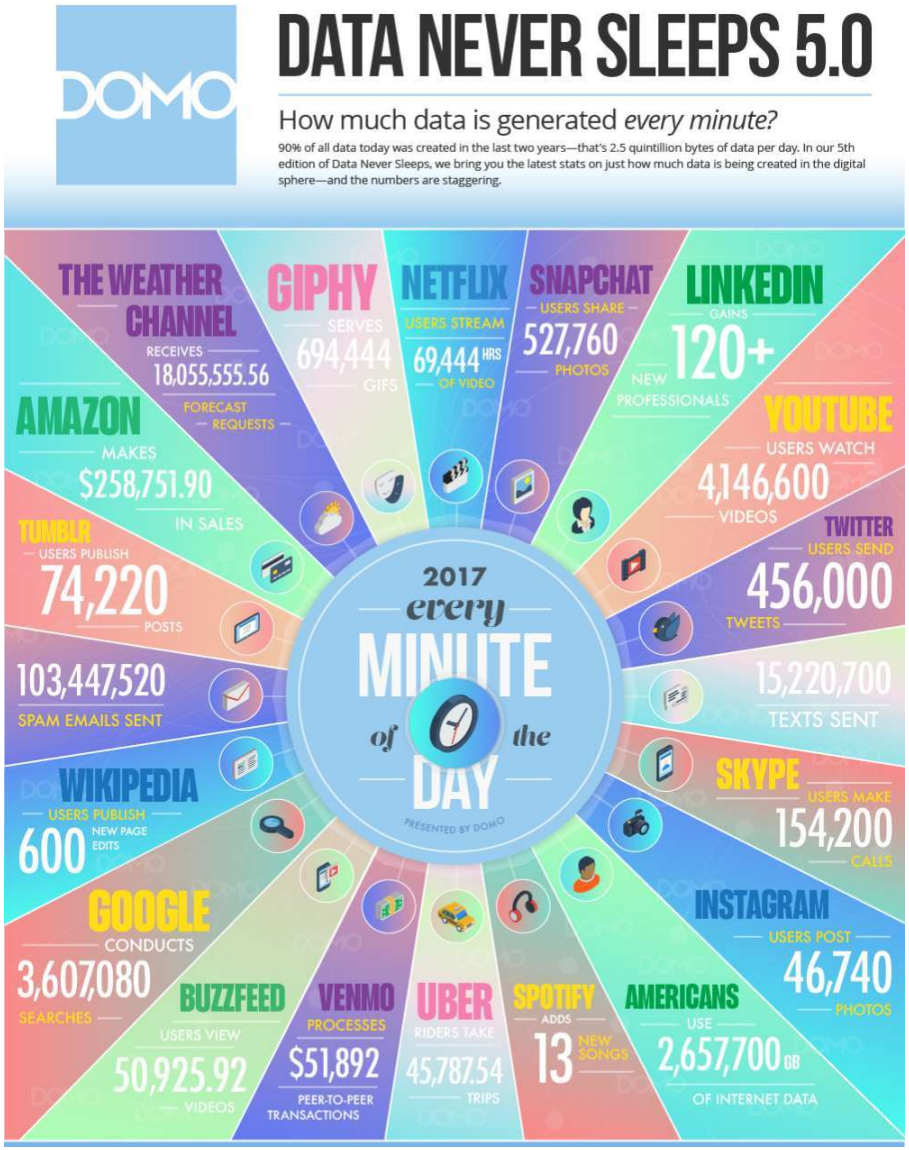
- Grade: 20%
- Loading textual data
- Building models
- Analysis
- Suggested programming language
  - Python 3
  - Scientific packages (e.g. scikit-learn)

Assignment policy

- Please read university regulations:
  - https://grad.uic.edu/university-regulations
- All assignment you turn must be done by you **alone**.
- The first violation will result in a failing grade for that assignment.
- The second will result in a failing grade for the course.
- **Late Submission Policy:**
  - 4% per hour
- **Grade dispute:**
  - Within 7 days of the receipt of the grade

Syllabus

- Computational linguistics:
  - Tokenization (TF-IDF)
- Supervised text mining (Logistic Regression, KNN)
  - Demographic classification
  - Opinion mining/sentiment
  - Word embedding (Deep Learning)
- Unsupervised text mining
  - Document/word clustering (K-means)
  - Dimensionality reduction (LSA, PCA)
  - Topic modeling (LDA)


DOMO
DATA NEVER SLEEPS 5.0
How much data is generated every minute?
90% of all data today was created in the last two years—that's 2.5 quintillion bytes of data per day. In our 5th edition of Data Never Sleeps, we bring you the latest stats on just how much data is being created in the digital sphere—and the numbers are staggering.
2017
every
MINUTE
of the
DAY
PRESENTED BY DOMO
THE WEATHER CHANNEL
RECEIVES
18,055,555.56
FORECAST REQUESTS
GIPHY
SERVES
694,444
GIFS
NETFLIX
USERS STREAM
69,444 HRS
OF VIDEO
SNAPCHAT
USERS SHARE
527,760
PHOTOS
LINKEDIN
GAINS
120+
NEW PROFESSIONALS
AMAZON
MAKES
$258,751.90
IN SALES
YOUTUBE
USERS WATCH
4,146,600
VIDEOS
TUMBLR
USERS PUBLISH
74,220
POSTS
TWITTER
USERS SEND
456,000
TWEETS
103,447,520
SPAM EMAILS SENT
15,220,700
TEXTS SENT
WIKIPEDIA
USERS PUBLISH
600
NEW PAGE EDITS
SKYPE
USERS MAKE
154,200
CALLS
GOOGLE
CONDUCTS
3,607,080
SEARCHES
INSTAGRAM
USERS POST
46,740
PHOTOS
BUZZFEED
USERS VIEW
50,925.92
VIDEOS
VENMO
PROCESSES
$51,892
PEER-TO-PEER TRANSACTIONS
UBER
RIDERS TAKE
45,787.54
TRIPS
SPOTIFY
ADDS
13
NEW SONGS
AMERICANS
USE
2,657,700 GB
OF INTERNET DATA

Data

- Structured
  - Pre-defined data model
- Unstructured (80% of all data)
  - Text
  - Audio
  - Photo
  - Video

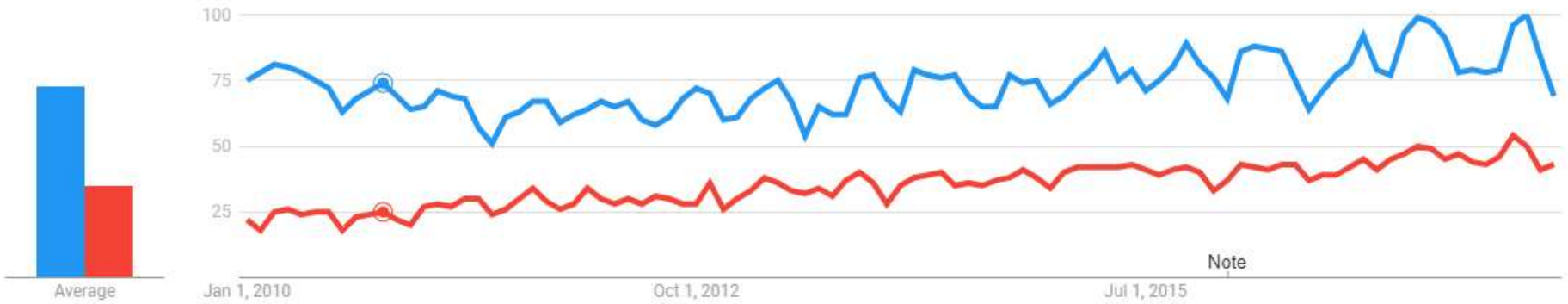
text mining
Search term

text analytics
Search term

+ Add comparison

Worldwide | 1/1/10 - 1/7/18 | All categories | Web Search

Interest over time



Text Mining

- Text mining refers to the process of deriving **high-quality** information from **text**.
- Text mining is the process of **discovering** and **extracting** knowledge from **unstructured** (textual) data.

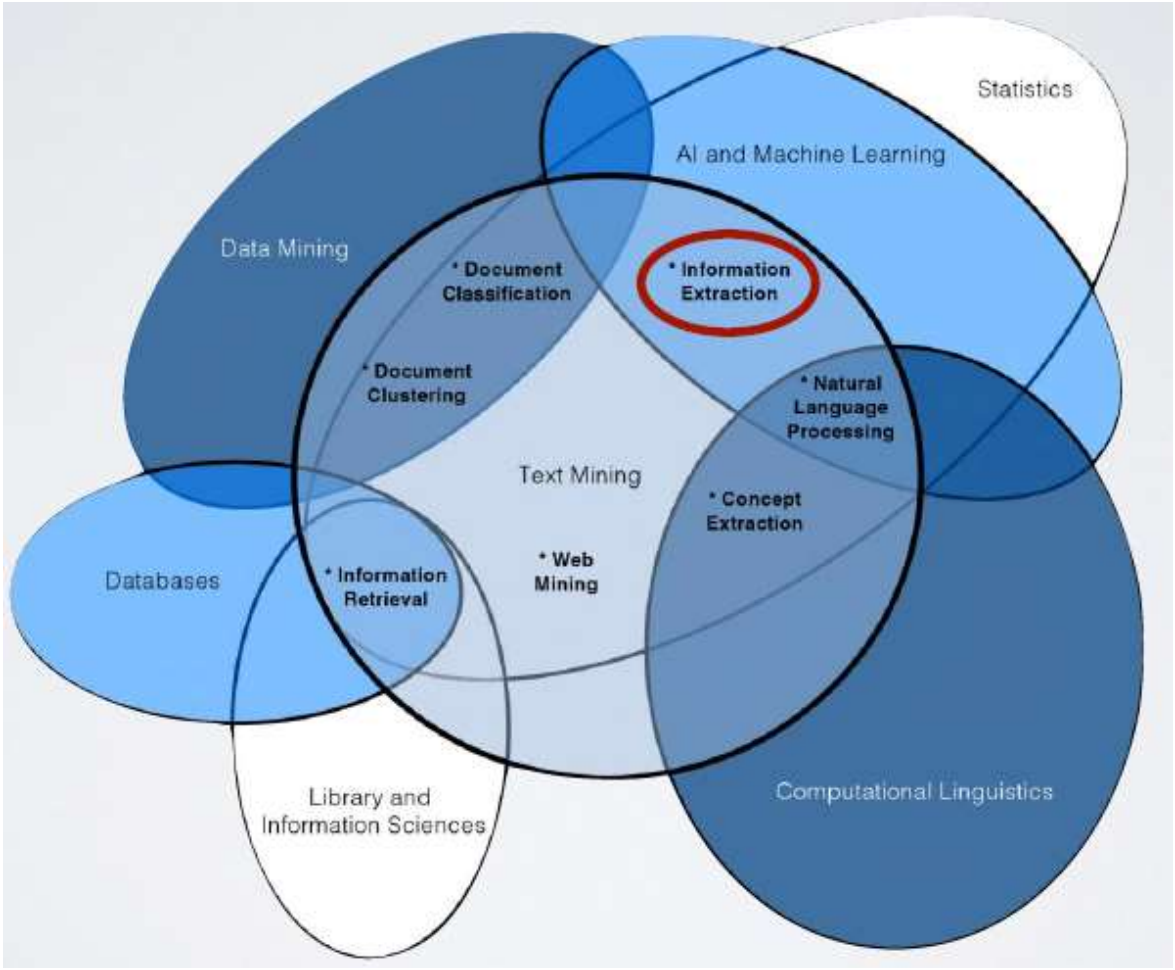
What is text mining?

Preprocessing

Deriving Patterns

Evaluation, Interpretation

Text Mining

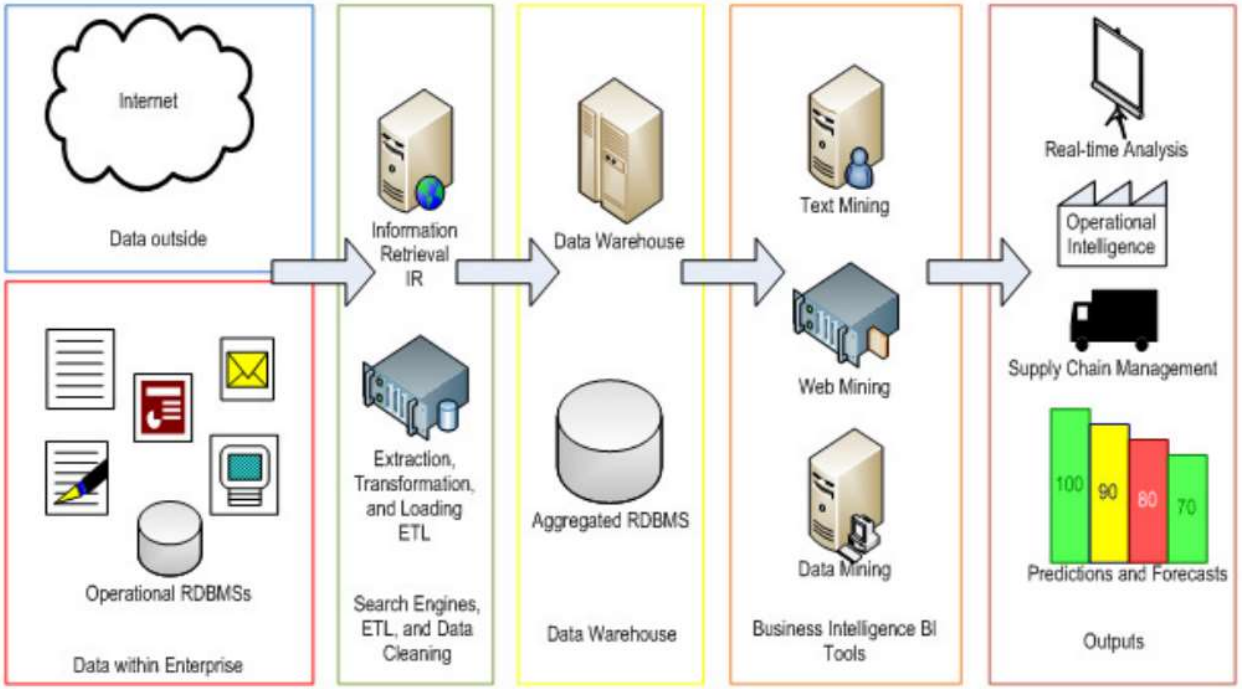

Statistics
AI and Machine Learning
Data Mining
* Document Classification
* Information Extraction
* Document Clustering
* Natural Language Processing
Text Mining
* Concept Extraction
Databases
* Information Retrieval
* Web Mining
Library and Information Sciences
Computational Linguistics

Generated by https://www.wordclouds.com/

Applications

- Business intelligence
- Sentiment analysis
- Social media analysis
- Personal news recommendation

Business Intelligence



International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.2, March 2013

Sentiment Analysis

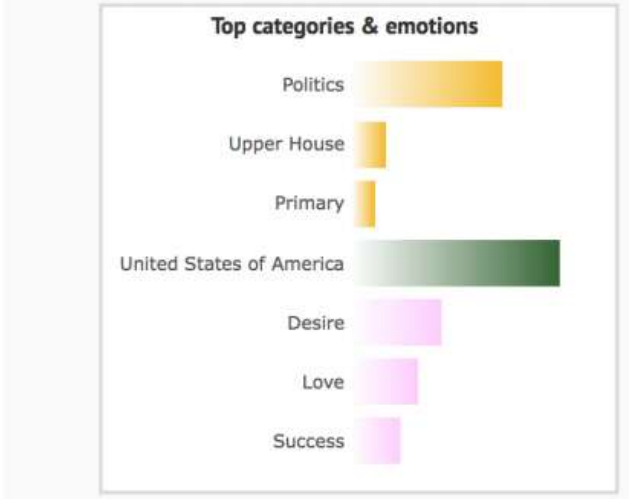
Expert IQ Report: Melania vs. Michelle – Divided Speeches

COMPARING MELANIA TRUMP AND MICHELLE OBAMA SPEECHES ANALYSIS

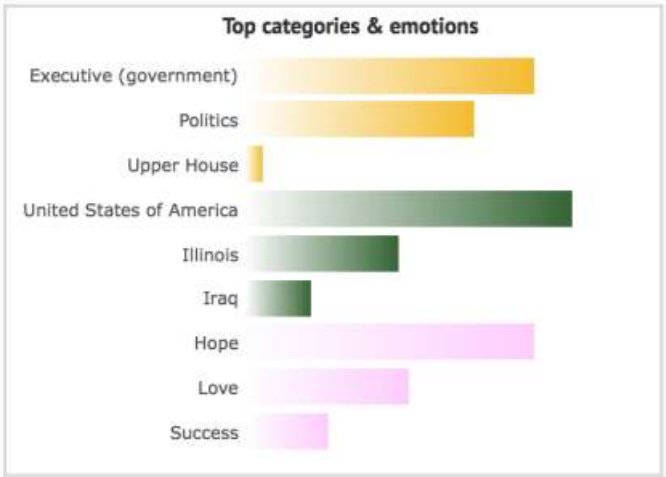
Expert System's Independent Text Analysis of Melania Trump and Michelle Obama's Speeches Confirms Strong Linguistic Differences

Source:CNN.com

Melania Trump

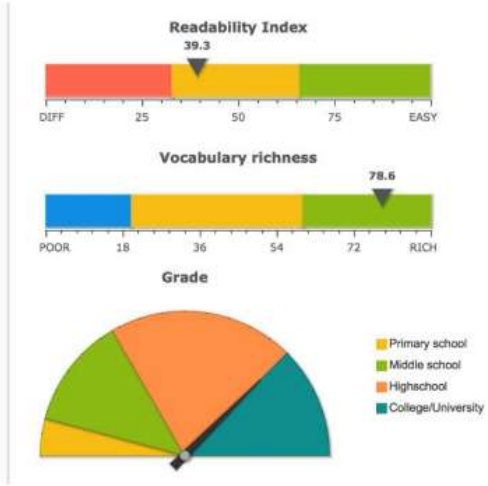


Michelle Obama

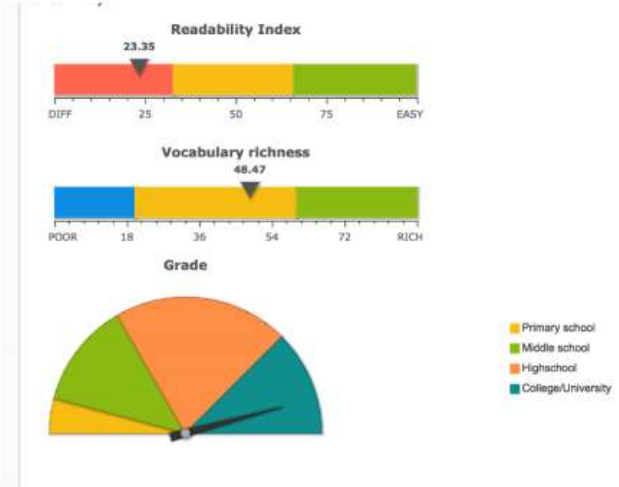


Sentiment Analysis

Melania Trump

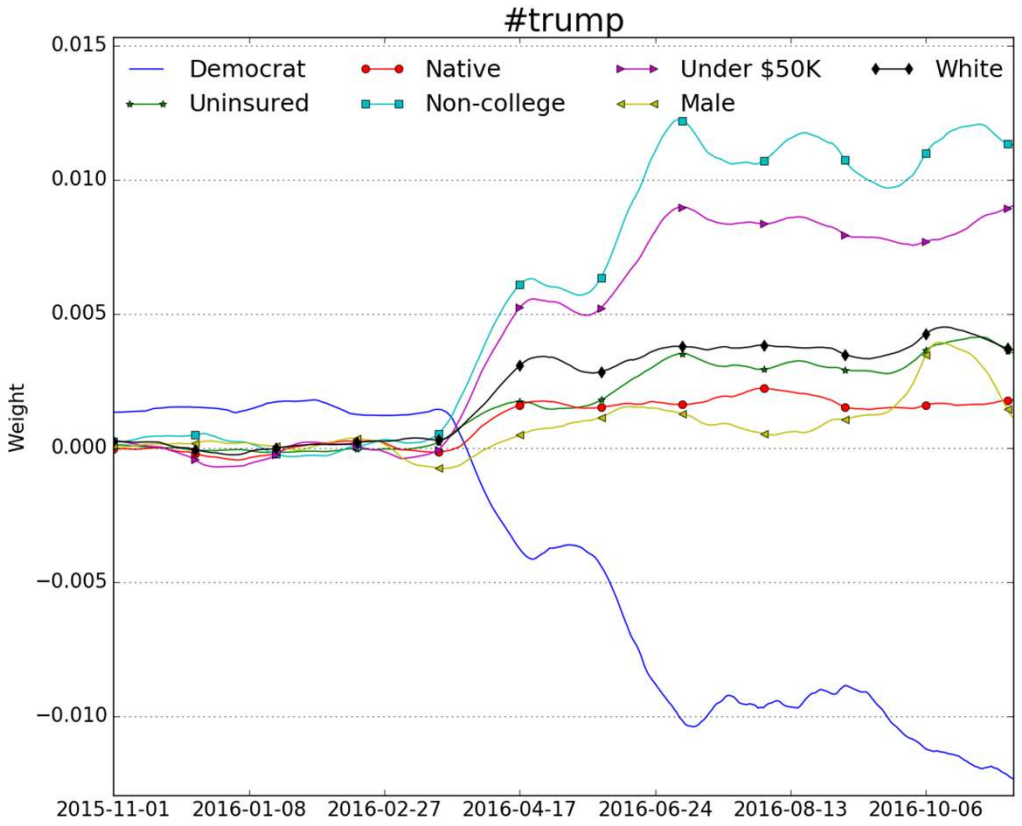


Michelle Obama



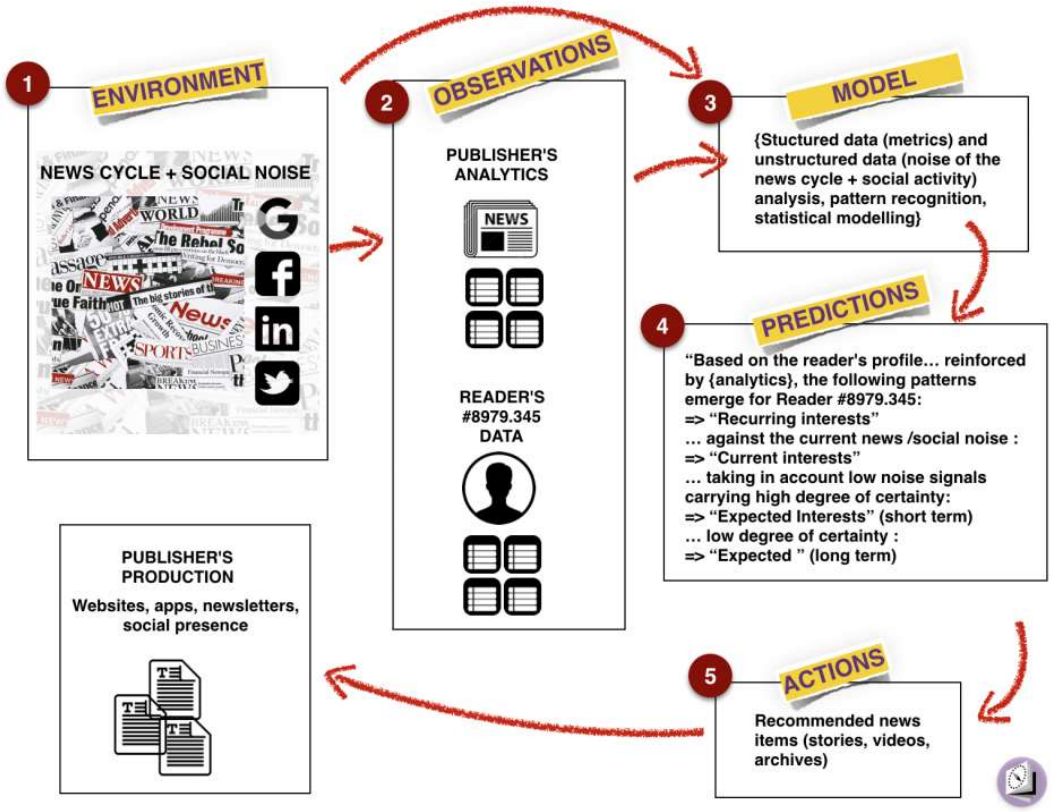
Lexical Analysis

Social Media Analysis



Mining the Demographics of Political Sentiment from Twitter Using Learning from Label Proportions, ICDM'17

News Recommendation



Google DeepMind, 2016

- Interpreted
- General purpose language
- Code readability
- Object Oriented
- Scientific libraries

Guido van Rossum, the creator of Python

Scientific Python

Scientific Packages

- Numpy: Array, Matrix, Linear Algebra, ...
- Scipy: Sparse Matrix, Sparse Linear Algebra, ...
- Scikit-learn: Machine Learning
- Tensorflow: Tensor operations, Deep Learning
- Keras: High-level Deep Learning
- NLTK: Natural Language Toolkit
- Matplotlib: Creating plots
- Networkx: Graphs, Networks

Python IDE

- **Spyder** (available free with Anaconda)
- Pycharm
- Netbeans
- Visual Studio

Jupyter

Jupyter Notebook

1 - Open a command prompt (shell).
2- Go to the course folder.
3- Type “jupyter notebook”.

Introduction to python

- Basic Python (lists, string, functions, conditions, ...)
- Numpy arrays
- Vector and Matrix operations and indexing
- Linear algebra
- Plots with matplotlib