Text Mining

Associate Professor
HUANG, Ke-Wei

Agenda

1. Method 4: Topic Modeling
   - Overview
   - Topic Modeling's High-Level Concepts
   - Formal Topic Modeling
   - Extensions
   - Examples of applications of Topic Modeling for Business Analytics

2. Method 5: Basics of Sentiment Analysis by Natural Language Processing
   - Examples of applications of Topic Modeling for Business Analytics

What is "Topic Modeling"?

- Given a (large) set of documents, Topic Modelling is an *unsupervised learning method that can tell you each document belongs to which topic*.
- The main weakness of supervised learning and text classification is sometimes we need to manually label the training set records => costly and errors may occur during the labelling process.
- Topic modelling is an unsupervised learning method in which you don't need to read/code any documents.

Inputs and Outputs of Topic Modeling

- Input: you need to decide (or try) different number of topics. The number of topics is an input to this kind of algorithm.

- Output 1: the algorithm will estimate a probability distribution of *K* topics for each document. That means the focal document, with X% chance, belongs to Topic *Y*.
- Output 2: The other by-product the algorithm will output is the dictionary of each of *K* topics => the probability distribution over words for each topic. => conditional on a given topic *Y*, what is the chance a word *Z* will appear.

Latent Dirichlet Allocation

- LDA (**Latent Dirichlet Allocation**) method is the most widely used algorithm.
- It is quite new and was invented in **2003**, see footnote for the original paper.
- Although it is the simplest method, the mathematical estimation is fairly complicated and is an evolving research topic for professors.
- The goal is to estimate 2 things: (1) a probability distribution over topics for each document, (2) dictionar**ies** = the probability distribution over words for each topic.

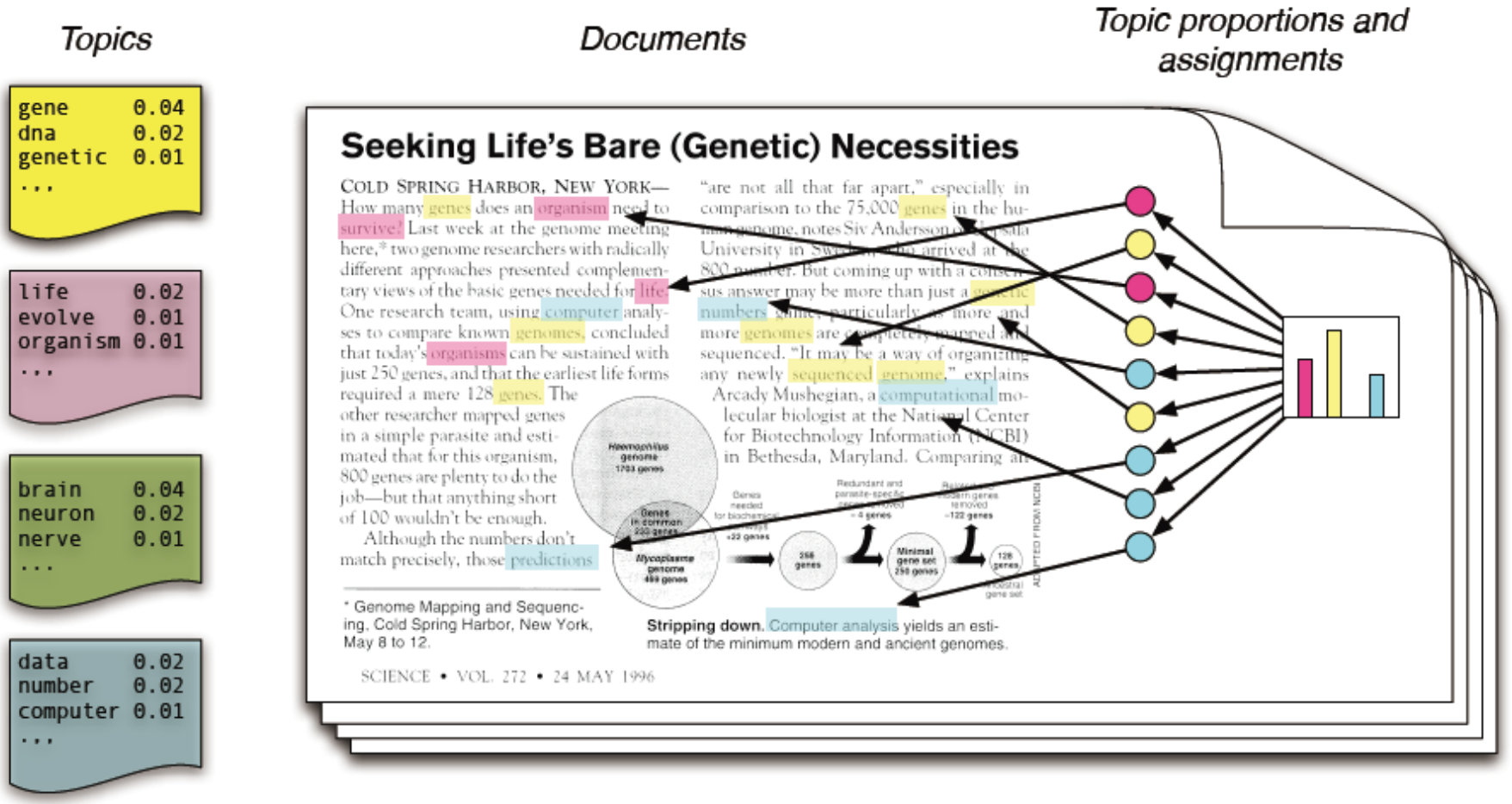


Figure 1: **The intuitions behind latent Dirichlet allocation.** We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.

Data Generating Process of Each Document

- The ASSUMPTIONs of this method are

(i) There are K topics.

(ii) Mathematically, each topic is defined as a probability distribution over all words.

(iii) **Each word** of a document is generated by randomly choose one topic, then randomly choose a word from the dictionary of that topic.

Given (ii) and (iii), now we can "create" a document.

For each word in the document,

- Step 1: we randomly pick one topic by distribution in (iii).
- Step 2: given a topic, we randomly pick one word from dictionary in (ii).

Topic Modeling

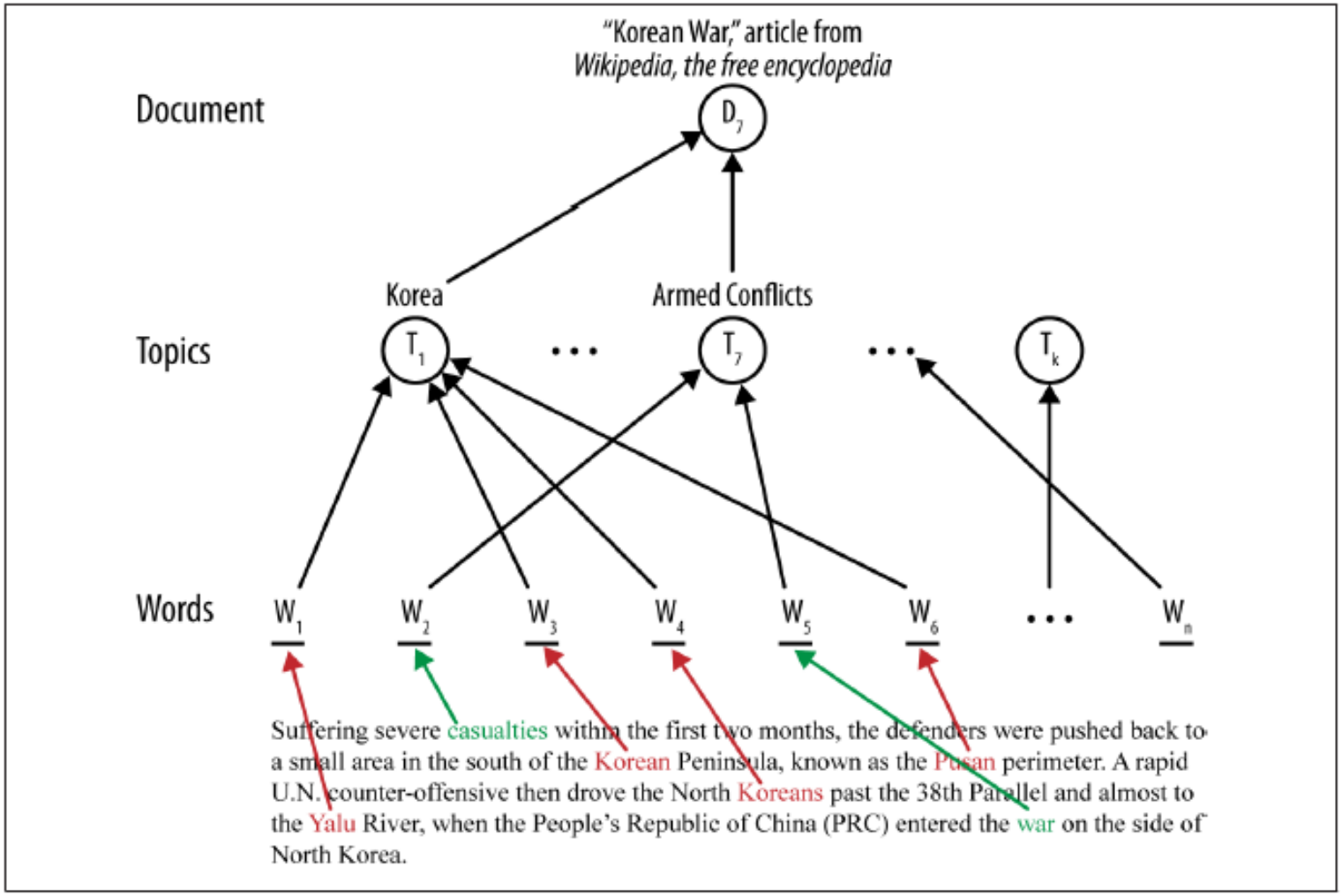
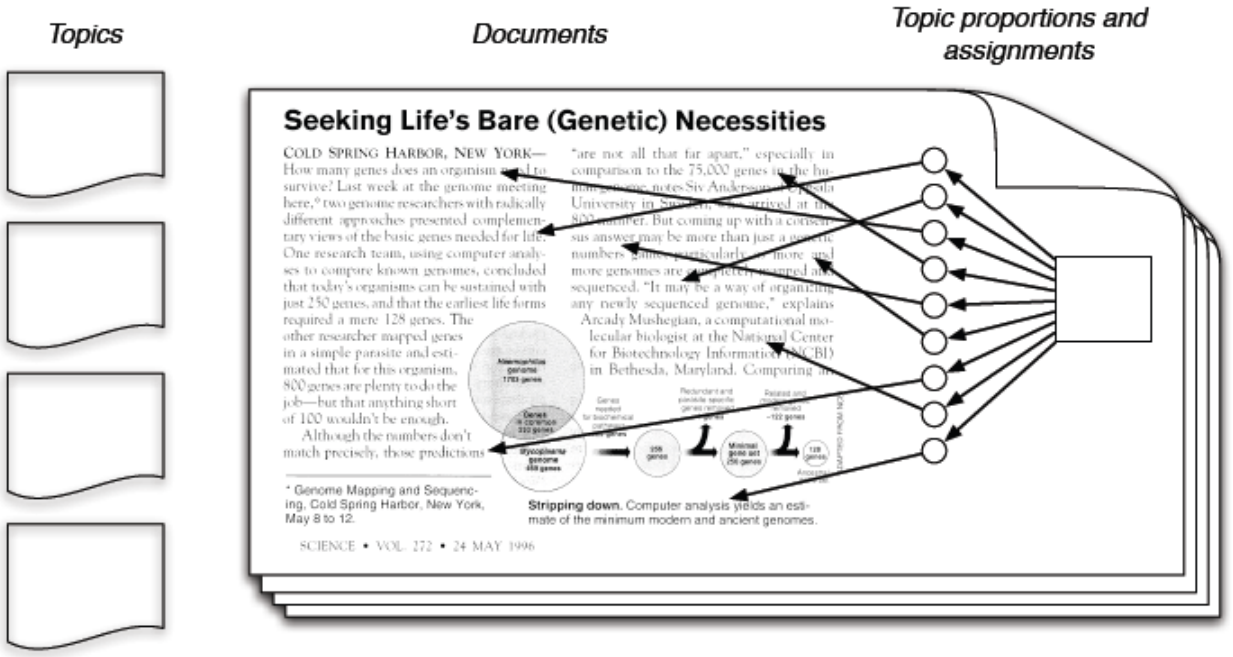


Figure 10-6. Modeling documents with a topic layer.

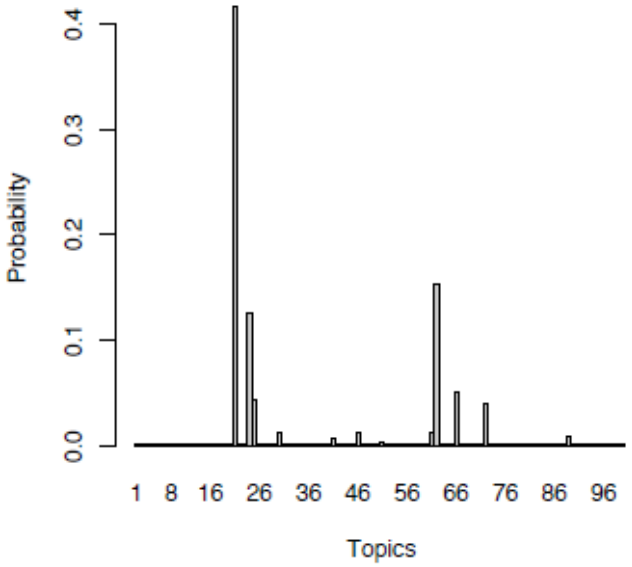
The posterior distribution



- In reality, we only observe the documents
- Our goal is to **infer** the underlying topic structure

In other words, given a large number of documents and input K, we need to infer the most likely hidden structures about (ii) and (iii) on the previous slide.

LDA



"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Figure 2: **Real inference with LDA.** We fit a 100-topic LDA model to 17,000 articles from the journal *Science*. At left is the inferred topic proportions for the example article in Figure 1. At right are the top 15 most frequent words from the most frequent topics found in this article.

LDA Formal Model

- The topics are β_1, β_2 ,… β_K, where each β_k is a probability distribution over the vocabulary.
- The topic probability distribution for the d^{th} document is denoted by θ_d, where $\theta_{d,k}$ is the topic proportion for topic *k* in document *d*.
- Beta and theta variables are the two objectives we try to estimate by LDA topic modeling method.
- The topic assignments for the d^{th} document are z_d, where $z_{d,n}$ is the topic assignment for the n^{th} word in document *d*.
- Finally, the observed words for document d are w_d, where $w_{d,n}$ is the n^{th} word in document d, which is an element from the vocabulary.

LDA Formal Model

With this notation, the generative process for LDA corresponds to the following joint distribution of the hidden and observed variables,

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right). \quad (1)$$

Only this is observable from data

This implies independence across words and documents

We now turn to the computational problem, computing the conditional distribution of the topic structure given the observed documents. (As we mentioned above, this is called the *posterior*.) Using our notation, the posterior is

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D} \mid w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}. \quad (2)$$

In general, this term cannot be computed. The core of academic research is to find methods to approximate this term in (2). Roughly speaking, existing methods are either by EM algorithms or modified Gibbs sampler to estimate this posterior probability. Both are beyond the scope of this module. Interested students can read the IVLE required reading and the other PDF uploaded to IVLE.

LDA Extensions (still growing now)

1. Advanced LDA can relax the bag of words assumption (order of words matters).
2. Another assumption of LDA is that the order of documents does not matter => Dynamic topic modeling to handle topics changing over time.
3. A third assumption (weakness) is that the number of topics is assumed known and fixed => Bayesian nonparametric topic model can estimate "the number of topics".
4. Bayesian nonparametric topic model can also estimate hierarchical topics.

LDA Extensions

5. You can seed keywords to affect the estimation of topics. See reference in the footnote.
6. In the standard LDA, topics are independent with each other. However, the occurrence of topics may be correlated. For example, for topic modeling on scientific articles, geology is more likely to co-occur with chemistry than finance. The solution is called "correlated topic model"
7. Incorporating meta-data: for example, if you try to predict the topics of academic papers and you also know the authors of each paper, then we can use this information ⇔ same author typically only writes on few topics.

Marketing Applications of Topics Modeling

1. “The Effect of Calorie Posting Regulation on Consumer Opinion” (2017).
   - 761,962 online reviews of restaurants posted over eight years.
   - Their model allows managers to specify prior topics of interest such as "health" for a calorie posting regulation.
   - New York City mandated that all chain restaurants post calorie information on their menus => (causes) there was a statistically small but significant increase in the proportion of discussion of the health topic.

Marketing Applications of Topics Modeling

2. “Mining marketing meaning from online chatter” (2014).
   - The sample of online user-generated content consists of rich data on product reviews across 15 firms in five markets over four years.
   - For vertically differentiated markets (e.g., mobile phones, computers), objective dimensions dominate and are similar across markets, heterogeneity is low across dimensions, and stability is high over time.
   - For horizontally differentiated markets (e.g., shoes, toys), subjective dimensions dominate but vary across markets, heterogeneity is high across dimensions, and stability is low over time.

Marketing Applications of Topics Modeling

3. Sentence-based text analysis for customer reviews (2016). (words in one sentence belongs to one topic.)

- Consumer feedback in the form of unstructured consumer reviews
- This paper proposes a new model for text analysis that makes use of the sentence structure contained in the reviews
- This paper shows that it leads to improved inference and prediction of consumer ratings relative to existing models using data from www.expedia.com and www.we8there.com.
- Sentence-based topics are found to be more distinguished and coherent than those identified from a word-based analysis.

Marketing Applications of Topics Modeling

4. “User profiling in customer-base analysis and behavioral targeting” 2016. (a unique application of topic modeling)

- User profile is a summary of a consumer's interests and preferences revealed through the consumer's online activity.
- This paper uncovers individual user profiles from online surfing data.
- Customer-base analysis and display advertising
- Search engines can effectively recover consumer behavioral profiles.
- Using simulation to demonstrate potential gains the proposed model may offer a firm if used in individual-level targeting of display ads.

Accounting & Finance Applications

1. Topic Modelling on Annual Report Risk Factors. "Simultaneously discovering and quantifying risk types from textual risk disclosures." (2014).
2. "The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation." (2017).
3. "Do fraudulent firms produce abnormal disclosure?." (2017).
4. "Marks of distinction: Framing and audience appreciation in the context of investment advice." (2015).
5. "Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud." (2018).
6. Many more papers…

Patenting Analysis Applications

Patent content analysis by topic modeling is a popular research topic in technology management literature.

1. "Topic based classification and pattern identification in patents.“ 2015
2. "Identification and monitoring of possible disruptive technologies by patent-development paths and topic modeling." 2016
3. "Firms' knowledge profiles: Mapping patent data with unsupervised learning." 2017

Economics Applications

Text mining and topic modeling has not been adopted in top economics journals. Only sparse papers published in 2nd or 3rd tier journals.

1. Hansen, Stephen, and Michael McMahon. "Shocking language: Understanding the macroeconomic effects of central bank communication." *Journal of International Economics* 99 (2016): S114-S133.
   - Text mining on FOMC meeting minutes or news announcement is one topic.
   - However, FOMC (central bank) related documents are difficult for human beings to understand its deeper meaning.
2. Azqueta-Gavaldón, Andrés. "Developing news-based Economic Policy Uncertainty index with unsupervised machine learning." *Economics Letters* 158 (2017): 47-50.

Other Business Applications

1. Predicting movies box office. "Early predictions of movie success: The who, what, and when of profitability." (2016):.
2. Automatically cluster companies. "Toward a Better Measure of Business Proximity: Topic Modeling for Industry Intelligence." (2016).
3. Online reviews for travel industry. "Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation." (2017).
4. Predicting online crowdfunding (e.g. Kickstarter.com, but this site uses China counterpart) success. "The determinants of crowdfunding success: A semantic text analytics approach." (2016).

Final Remarks

- Topic modeling alone may not give you very good results; you can mix and match topic modeling with text classification or other methods to achieve your analytics objectives.

Programming Resources

1. Original Author: David Blei's page. (He moved from Princeton to Columbia)
   http://www.cs.columbia.edu/~blei/
2. Stanford has a very strong NLP group with many resources available online. They publish a topic modelling toolbox.
   Stanford Topic Modeling Toolbox
   https://nlp.stanford.edu/software/tmt/tmt-0.4/

Agenda

1. Method 4: Topic Modeling
   - Overview
   - Topic Modeling's High-Level Concepts
   - Formal Topic Modeling
   - Extensions
   - Examples of applications of Topic Modeling for Business Analytics

2. Method 5: Basics of Sentiment Analysis by Natural Language Processing
   - Examples of applications of Topic Modeling for Business Analytics

2. Natural Language Processing

- Natural language processing is a sophisticated and broad research discipline that involves studying algorithms for computer to process, understand, or translate a large amount of textual information.
- There is still very limited NLP applications in social science research due to the complexity in understanding how NLP algorithms work.
- Deep learning methods can be applied to NLP problems to achieve state-of-the-art results. It is very hot in CS to study NLP problems with deep learning methods.

NLP research areas related to Business Analytics

- Name-entity recognition: helps you identify person or organization names in a document.
- String fuzzy matching for correcting typo or matching two strings (for merging datasets by names).
- Information Extraction: extracting specialized information that you are looking for from a large amount of documents.
- Text summarizing: computers can summarize a long document without losing much meaning.
- Chat bot is a hot commercial application.
- Q&A system (behind Amazon Echo, Google Home) and Knowledge graph is one fruitful area of NLP research.

Natural Language Processing Tool

- The science behind NLP is complicated but to use software packages for analytics is easier.
- Online demo http://nlp.stanford.edu:8080/parser/

- A new sentiment NLP parser for your future reference. http://nlp.stanford.edu/sentiment/code.html

Example 1 input:

Having stayed in HV, the condo is cosy and home welcoming. The environment is quiet and beautiful and the condo interior is truly 'resort-like'. Very relaxing coming back home everyday....

Example: Output

- See footnotes for full output.
- We can focus on nsubj(,) and amod(,) to extract the adj. and noun pairs.
  - nsubj(cosy-9, condo-7)
  - nsubj(quiet-4, environment-2)
  - nsubj(resort-like-14, interior-10)
  - nsubj(gem-11, It-8)
- Example 2: Huge development with complete and impressive facilities even though it is a bit aged.
  - amod(development-2, Huge-1)
  - nsubj(bit-13, development-2)
  - amod(facilities-7, complete-4)
  - cc(complete-4, and-5) conj(complete-4, impressive-6)
  - nsubj(bit-13, it-10)

Example: Output

- Example 3: Huge development with complete and impressive facilities even though **it is aged**.
  - amod(development-2, Huge-1)
  - nsubjpass(aged-12, development-2)
  - amod(facilities-7, complete-4)
  - cc(complete-4, and-5) conj(complete-4, impressive-6)
  - nsubjpass(aged-12, it-10)
- Example 4: Huge development with complete and impressive facilities even though **it is not aged**.
  - amod(development-2, Huge-1)
  - nsubjpass(aged-13, development-2)
  - amod(facilities-7, complete-4)
  - cc(complete-4, and-5) conj(complete-4, impressive-6)
  - nsubjpass(aged-13, it-10) **neg(aged-13, not-12)**

Simple NLP Sentiment Analysis

Step 1:

Apply existing NLP parser packages (such as the Stanford one) to your corpus.

Step 2:

Given the tags, you know the nouns. Identify the keywords of interests (your company or competitors' name, brand name, product name, …etc.)

Step 3a: Given the pair function (called Universal dependencies), you can identify the keywords that associated with your focal words (such as your brand/product names). Then apply dictionary approach to extract the sentiment. You can also include negate in this case.

Step 3b: other methods are possible, for example, trace along the tree to find associated adj. keywords.

Simple NLP Sentiment Analysis Example

- "Show me the money!: deriving the pricing power of product features by mining consumer reviews." KDD 2007
- Step 1: Applying NLP POS-tagging to identify nouns that represent product features in Amazon consumer reviews.
- Step 2: they use other methods to identify important product features.
- Step 3: using Stanford parser to identify the adj. that is used to describe each features.
- Step 4: the authors use the extracted sentiment to run a regression on sales rank to find out which product feature in review comment is more influential.

For your projects: Data Sources

1. Kaggle.com for data competition
2. Crawl from the Internet websites (all kinds of reviews)
3. UC Irvine KDD archives https://kdd.ics.uci.edu/
4. KDD Cup datasets (large and complicated)
5. Data challenge hosted by large companies: e.g., Yelp for online consumer reviews, Foursquare, LinkedIn, and several other firms hosted data challenges in the past. You can even try active ongoing ones.
6. NUS subscribed financial/accounting/economics datasets (Compustat, CRSP, I/B/E/S, …etc.)
7. NUS news database: Factiva and LexisNexis
8. Google Trends can give you search volume of keywords as one feature or DV for prediction.

Accounting & Finance Documents

1. Financial news
2. Online forums about investment
3. Other SEC (USA) filings
4. Annual Reports or Quarterly Reports from SEC
5. Earnings conference script
6. Analysts reports
7. FOMC meeting minutes
8. FB, Twitter tweets and other social media messages
9. Company/CEO reviews at Glassdoor
10. Other unique data sources…

For your projects: Text Analytics for Marketing

- Predicting the impact of consumer reviews on sales using the valence of sentences (Berger et al. 2010)
- Determining the relative importance of reviews in comparison to own experience in the learning process of consumers about products (Zhao et al. 2013)
- Analyzing the change in conversion rates as a result of changes in affective content and linguistic style of online reviews (Ludwig et al. 2013)
- Predicting the sales of a product based on review content and sentiment (Godes and Mayzlin 2004, Dellarocas et al. 2007, Ghose et al. 2012)
- Eliciting product attributes and consumers preferences for attributes (Lee and Bradlow 2011, Archak et al. 2011)
- Deriving market structure (Netzer et al. 2012, Lee and Bradlow 2011)

For your projects: Remarks about Text Classification

- First, you need to create the label of a small subset of documents because this is supervised learning.
- Typically, you need to read documents to code the label of the documents. It may be time consuming but you also have the flexibility in coding documents to create any label that you need.
- Wrong example/application. You have online review ratings. Then you train a text classification model to predict rating from textual reviews. Next, you use the predicted ratings in your main model to predict something else. This is wrong because you should just use the actual review ratings in your main model. Text classification is correct only when you code textual reviews to create a categorical variable that does not exist in your raw data set.