Strategic Data Science (SDS)

Text Data

Karl Ho

School of Economic, Political and Policy Sciences

University of Texas at Dallas

What is Text Data?

What is Text Data?

Text data refer to any documents or corpus in text forms.

What is Text Data?

Text data refer to any documents or corpus in text forms.

- Structured data refers to text with a high degree of organization, such that inclusion in a relational database is seamless and readily searchable.

What is Text Data?

Text data refer to any documents or corpus in text forms.

- Structured data refers to text with a high degree of organization, such that inclusion in a relational database is seamless and readily searchable.
- Unstructured data do not have a pre-defined data model or is not organized in a pre-defined manner.

What is Text Data?

The TEXT data type stores any kind of text data. It can contain both single-byte and multibyte characters that the locale supports.

- IBM

Terminology

Terminology

Corpus (plural Corpora) - A set of multiple similar documents is called a corpus.

Terminology

Corpus (plural Corpora) - A set of multiple similar documents is called a corpus.

Tokenization - The first step in processing text is deciding what terms and phrases are meaningful. Tokenization separates sentences and terms from each other.

Terminology

Terminology

N-grams -

a contiguous sequence of n items from a given sequence of text or speech.

These phrases often appear together in fixed patterns such as “systems of innovation,” “cease and desist,” or “commander in chief.”   These combinations of phrases are also called collocations, as their overall meaning is more than the sum of their parts.

Terminology

Terminology

Stop words are a category of words that have limited semantic meaning regardless of the document contents. Such words can be prepositions, articles, common nouns, etc.

Terminology

Stop words are a category of words that have limited semantic meaning regardless of the document contents. Such words can be prepositions, articles, common nouns, etc.

For example, the words “the”, “to” and “of”, which could account for more than 10 percent of the text.

Terminology - stop words

Terminology - stop words

Example:

http://www.analytictech.com/mb021/mlk.htm

Terminology - stop words

Example:

http://www.analytictech.com/mb021/mlk.htm

Can you identify other stop words manually?

Computer-aided Text Analysis

Computer-aided Text Analysis

- Searches and information retrieval

Computer-aided Text Analysis

- Searches and information retrieval
- Clustering and text categorization

Computer-aided Text Analysis

- Searches and information retrieval
- Clustering and text categorization
- Text summarization

Computer-aided Text Analysis

- Searches and information retrieval
- Clustering and text categorization
- Text summarization
- Machine translation

Text Mining

Text Mining

Text mining refers to the practice of extracting useful analytic information from corpora of text.

- Saltz and Stanton 2018

Text Mining vs. NLP

Text Mining vs. NLP

Text mining looks for patterns in large data sets. Natural Language Processing (NLP) is more sophisticated on studying how machines can be programmed to digest and make sense of human language.

Approaches in Text Mining

Approaches in Text Mining

- Bag of words

Approaches in Text Mining

- Bag of words
- Topic modeling

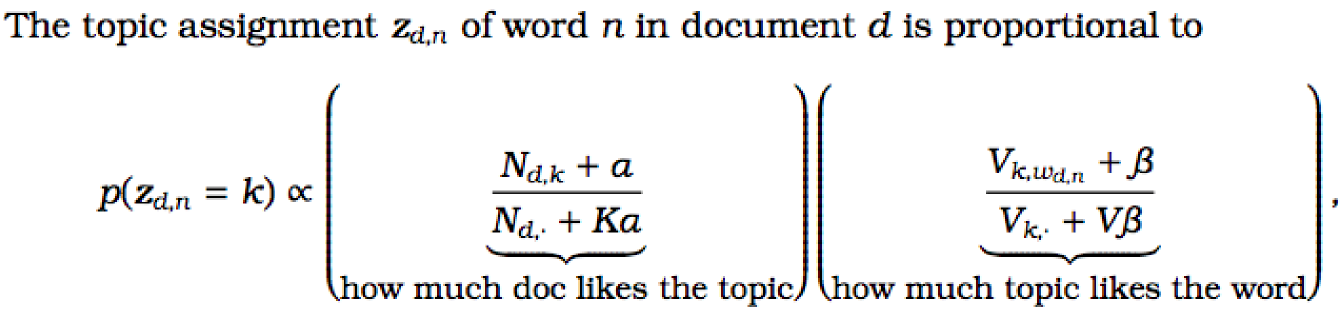
Approaches in Text Mining

- Bag of words
- Topic modeling
- automated classification and clustering

Approaches in Text Mining

- Bag of words
- Topic modeling
- automated classification and clustering
- dimensionality reduction

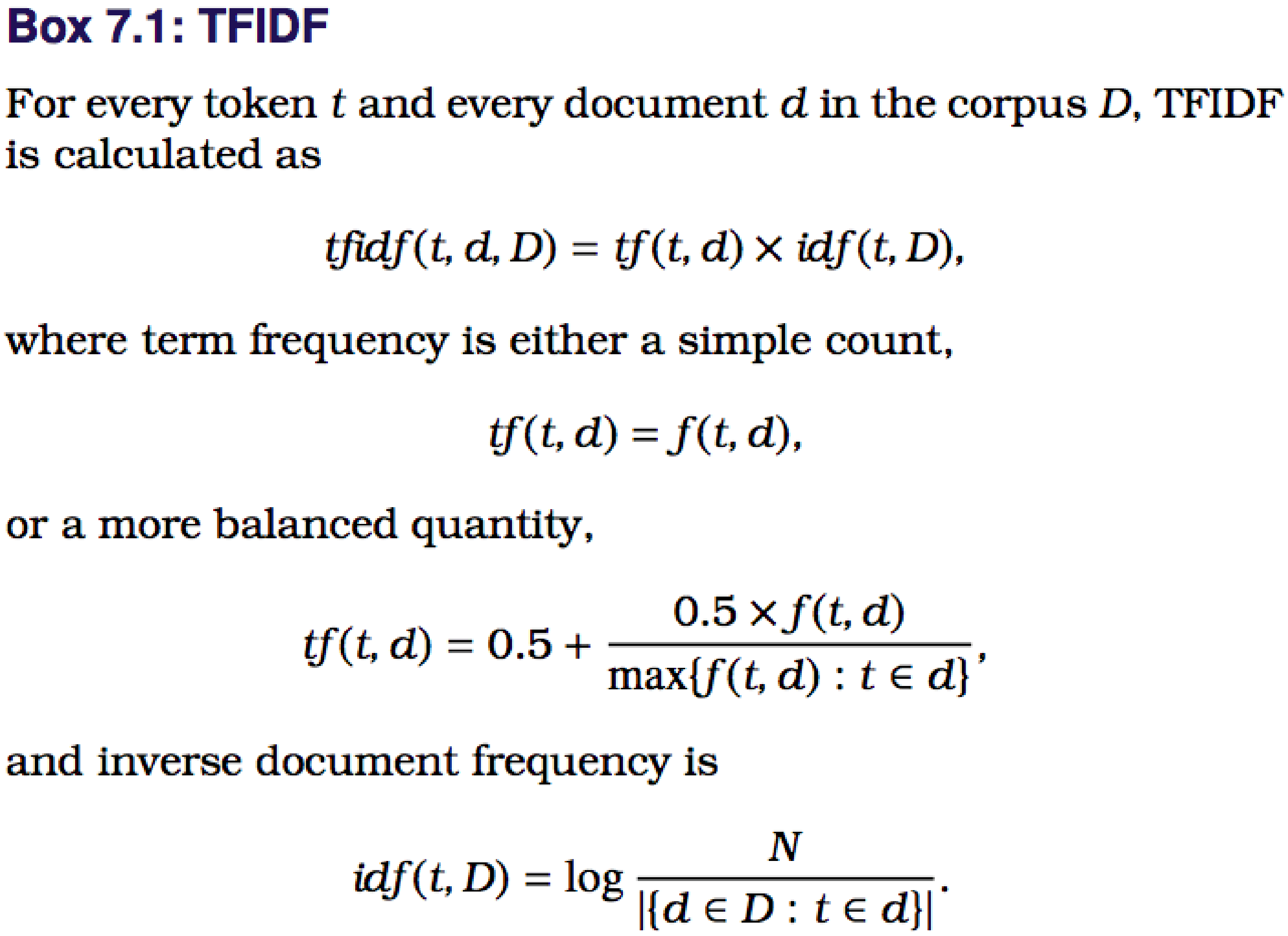
Gibbs Sampling for Topic Models



The topic assignment $z_{d,n}$ of word n in document d is proportional to

$$p(z_{d,n} = k) \propto \left(\underbrace{\frac{N_{d,k} + \alpha}{N_{d,\cdot} + K\alpha}}_{\text{how much doc likes the topic}} \right) \left(\underbrace{\frac{V_{k,w_{d,n}} + \beta}{V_{k,\cdot} + V\beta}}_{\text{how much topic likes the word}} \right),$$

Term frequency–inverse document frequency (TFIDF)



Box 7.1: TFIDF

For every token t and every document d in the corpus D, TFIDF is calculated as

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D),$$

where term frequency is either a simple count,

$$tf(t, d) = f(t, d),$$

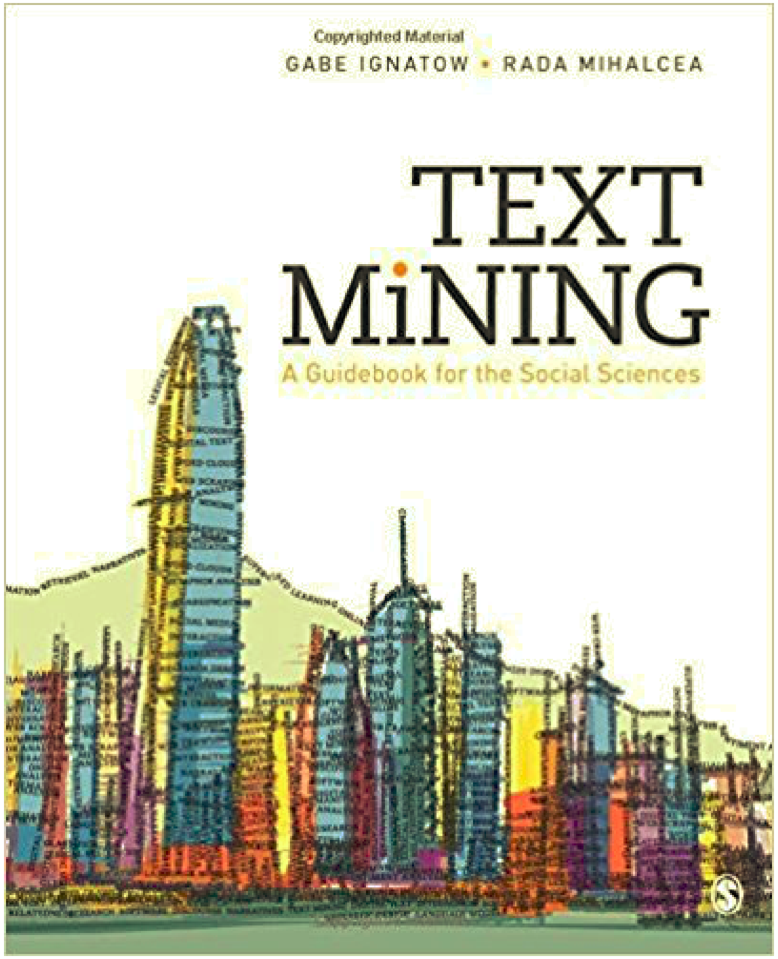
or a more balanced quantity,

$$tf(t, d) = 0.5 + \frac{0.5 \times f(t, d)}{\max\{f(t, d) : t \in d\}},$$

and inverse document frequency is

$$idf(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}.$$

Further Reading


Copyrighted Material
GABE IGNATOW • RADA MIHALCEA
TEXT
MiNING
A Guidebook for the Social Sciences

Aggarwal, C.C. and Zhai, C. eds., 2012. *Mining text data*. Springer Science & Business Media.

Ignatow, G. and Mihalcea, R., 2016. *Text Mining: A Guidebook for the Social Sciences.* Sage Publications.

Workshop in RStudio

Workshop in RStudio

Objective:

- Access and analyze unstructured data
- Be familiar with word clouds
- Apply R packages to do basic text mining