Graph-Based Term Weighting for Topic Modeling

Giannis Bekoulis and François Rousseau

SoMeRis workshop, ICDM 2016

January 6, 2017

Abstract

Challenge

- Topic models for **long** documents

Introduce

- More realistic assumption into topic models
- **Graph-of-words** representation of textual documents
- Alternative weighting mechanism for topic models based on graph theory

Evaluation

- Single-label multi-class text categorization

Outline

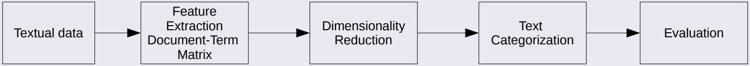
Outline

Introduction

- Online social media and networking platforms produce a vast amount of textual data
- Analyze and extract useful information from textual data is a crucial task
- Model the large data collections - **topic models**
- **Text categorization** assign a document to a set of predefined categories
- Applications:
  - Opinion mining for risk assessment and management
  - Email filtering
  - Spam detection
  - ...

Text categorization with dimensionality reduction

Pipeline



Pipeline: Topic Modeling, Text/Graph Mining

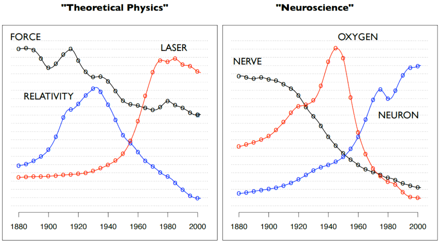
- **Feature extraction** from textual information
- Use topic models for **dimensionality reduction**
- Evaluate through **text categorization**

Topic Modeling

- **Mathematical** framework
- Model the **underlying structure** from a collection of documents
- **Soft-clustering** algorithms

Baseline methods

- LSI (Latent Semantic Indexing)
- LDA (Latent Dirichlet Allocation)



"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

- Used by humanists, social scientists, computer scientists to analyze big text corpora
- Applied for image categorization, image topic extraction and analyze topic evolution
- Visualize themes from document collections

Outline

1. Introduction
2. Document Representation
3. Proposed weighting scheme
4. Conclusions & Future work

Term independence: Bag-of-words representation

Bag-of-Words

In the **bag-of-words** model, a sentence or an entire document is represented as the multi-set of its words/terms, disregarding grammar and even word order.

Bag-of-Words is the traditional way of representing a document in the Vector Space Model

- Raw frequency of a term in document (TF)

Transformation

- Learn vocabulary from the training set
- Transform documents into document term matrix

Contributions of this work

Graph-based term weighting schemes for Topic Modeling

- Propose a simple graph-based representation of documents for topic modeling
- Derive novel term weighting schemes, that go beyond single term frequency

Exploration of model's parameter space and experimental evaluation

- We discuss how to construct the graph
- We examine the performance of the different proposed weighting criteria using standard document collections

Outline

Motivation

- The terms are **not independent**
- Otherwise, the documents would be unreadable

Challenges

- Find another more elegant way to represent raw documents questioning bag-of-words's independence assumption
- Capture the relationships between the terms
- Introduce more **realistic feature weights**

Proposed Approach

- Graph-of-words
- Already applied in other data analytics tasks (e.g., IR [**?**], text classification [**?**])

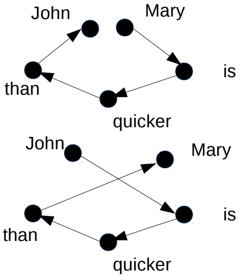
Alternative representation of a document that captures the relationships between the terms using graph of terms

$G(V, E)$

- Nodes correspond to the terms t of the document
- Edges capture co-occurence relations between terms within a fixed-size sliding window of size w

2 equivalent BoW sentences

- Mary is quicker than John
- John is quicker than Mary



Directed vs. undirected graph

- Directed graphs are able to preserve actual **flow of a text**
- In undirected ones, an edge captures **co-occurrence** of two terms whatever the respective order between them is

Weighted vs. unweighted graph

- Weighted: the higher the number of **co-occurences** of two terms in the document, the higher the weight of the **corresponding edge**
- Unweighted

Size w of the sliding window

- We add edges between the terms of the document that co-occur within a sliding window of size w
- $w = 3$ performed well in Topic Modeling

- **Degree:** in an **undirected** graph captures the **number of neighbors** that each node has.
- **In-degree:** in a **directed** graph captures only the **incoming** edges
- **Out-degree:** in a **directed** graph captures only the **outgoing** edges
- **Weighted**: is an extension for **weighted undirected** graphs

Example

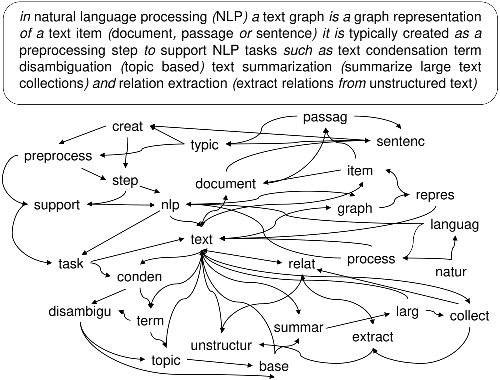


Figure: Directed graph-of-words representation of the text in the upper box. Nodes correspond to unique terms and edges to co-occurrences within a window of size 3.

Evaluation setup

Datasets

- 20ng: 18,821 newsgroup documents of 20 categories
  - # of train docs: 11,293;# of test docs: 7,528;
- Reuters R8: 8 most frequent categories of Reuters-21578
  - # of train docs: 5,485;# of test docs: 2,189;
- Reuters R52: 52 most frequent categories of Reuters-21578
  - # of train docs: 6,532;# of test docs: 2,568;
- BBC news:2,225 documents from the BBC news website
  - # of train docs: 1,780;# of test docs: 445;

Evaluation

- Evaluate the performance of topic models using TF and graph-based methods
- Text classification on **long** text documents

	method \ dataset	20ng Acc	20ng F_1	R8 Acc	R8 F_1	R52 Acc	R52 F_1	BBC Acc	BBC F_1
LSI	TF (baseline)	0.7125	0.7032	0.9246	0.7582	0.8298	0.2696	0.8966	0.8953
LSI	TW_u (degree)	0.7055	0.6982	0.9223	**0.7718**	**0.8462***	**0.3944**	0.9326*	0.9331
LSI	TW_{uin} (in degree)	**0.7614***	**0.7503**	0.9278	0.7331	0.8166	0.1997	**0.9371***	**0.9353**
LSI	TW_{uout} (out degree)	0.7398*	0.7306	**0.9333***	0.7699	0.8368	0.2818	**0.9371***	0.9351
LSI	TW_w (weighted)	0.6869	0.6779	0.9141	0.7511	0.7960	0.3380	0.9191*	0.9145
LDA	TF (baseline)	0.7194	0.7031	0.7958	0.3594	0.6783	0.0504	0.8315	0.8267
LDA	TW_u (degree)	0.7388*	0.7248	0.7985	0.3778	0.6807	0.0551	0.8584	0.8583
LDA	TW_{uin} (in degree)	0.7325*	0.7229	0.7775	0.3085	0.6632	0.0439	0.8494	0.8461
LDA	TW_{uout} (out degree)	0.7198	0.7065	0.7967	0.3909	0.6791	0.0553	**0.8876***	**0.8852**
LDA	TW_w (weighted)	**0.7392***	**0.7272**	**0.8164***	**0.4327**	**0.6967***	**0.0673**	0.8607*	0.8599

Outline

Conclusions

Conclusions

- **graph-based features** are more **discriminative** for topic models in the case of **long documents**
- unweighted node degrees for LSI
- weighted node degrees for LDA
- use graph-based features to extract themes/topics from a collection of documents

Future work

Consider a **graph-normalization** scheme over the whole collection similar to IDF

References I

Blanco, R. and Lioma, C. (2012). Graph-based term weighting for information retrieval. *Information Retrieval*, 15(1):54–92.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, HLT '11, pages 142–150.

Malliaros, F. D. and Skianis, K. (2015). Graph-based term weighting for text categorization. In *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1473–1479.

Mihalcea, R. and Tarau, P. (2004). TextRank: Bringing Order into Texts. In *Proceedings of the 9th Conference on Empirical Methods in Natural Language Processing*, EMNLP '04, pages 404–411.

Rousseau, F., Kiagias, E., and Vazirgiannis, M. (2015). Text Categorization as a Graph Classification Problem. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, volume 1 of *ACL-IJCNLP '15*, pages 1702–1712.

Rousseau, F. and Vazirgiannis, M. (2013). Graph-of-word and tw-idf: New approach to ad hoc ir. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 59–68.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.