Understanding Semantic Change of Words Over Centuries

UFABC

O TIME

Jairo Modelagem

Mayza Análise e apresentação

Rafael Análise e apresentação

Rodrigo Coleta e modelagem

Understanding Semantic Change of Words Over Centuries

Derry Tanti Wijaya
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213
dwijaya@andrew.cmu.edu

Reyyan Yeniterzi
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213
reyyan@cs.cmu.edu

ABSTRACT

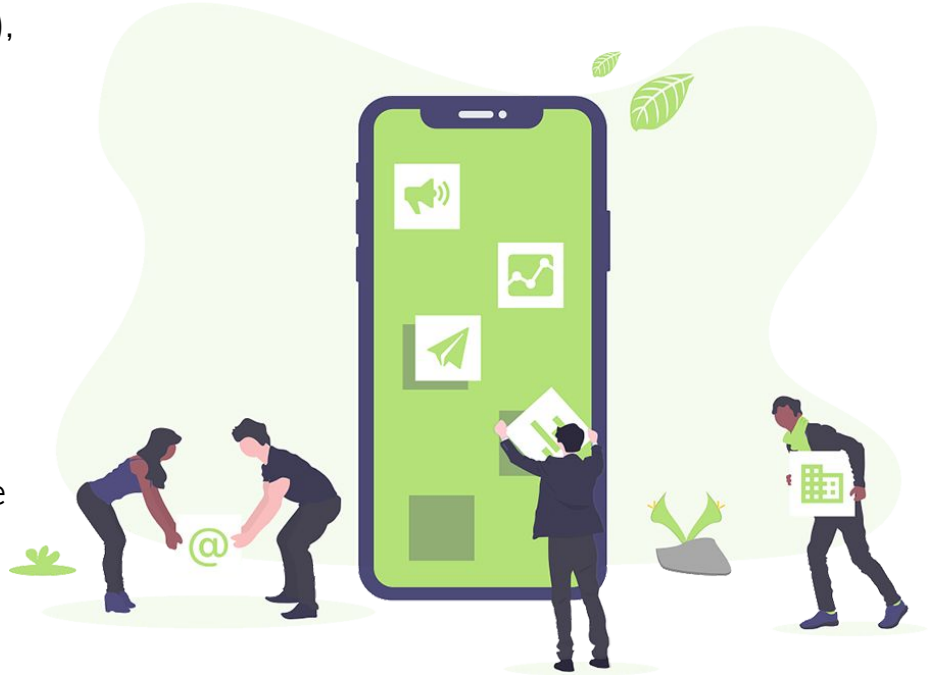
In this paper, we propose to model and analyze changes that occur to an entity in terms of changes in the words that co-occur with the entity over time. We propose to do an in-depth analysis of how this co-occurrence changes over time, how the change influences the state (semantic, role) of the entity, and how the change may correspond to events occurring in the same period of time. We propose to identify clusters of topics surrounding the entity over time using Topics-Over-Time (TOT) and k-means clustering. We conduct this analysis on Google Books Ngram dataset. We show how clustering words that co-occur with an entity of interest in 5-grams can shed some lights to the nature of change that occurs to the entity and identify the period for which the change occurs. We find that the period identified by our model precisely coincides with events in the same period that correspond to the change that occurs.

contexts, the meanings of the words change gradually, often to the point that the new meaning is radically different from the original usage. For example, awful [1] originally meant 'awe-inspiring, filling someone with deep awe', as in *the awful majesty of the Creator*. At some point it becomes something extremely bad, as in an awfully bad performance, but now the intensity of the expression has lessened and the word is now used informally to just mean 'very bad', as in *an awful mess*. Some words also change semantically, not in their original meanings but change in a way that they acquire additional meanings or are used to refer to other named entities over time. For example, mouse is used originally to refer to small long-tailed animal but it is now also used to refer to a device used to control cursor movement.

Automatically identifying changes to an entity over time is beneficial to many natural language applications. For example, for a macro-reader [2] that gathers 'background/commonsense' facts about entities from a large collection of input

Sobre o artigo:

- 2011 - Conference on Information and Knowledge Management da ACM (ACM - CIKM), Escócia
- DETECT - DETecting and Exploiting Cultural diversiTy on the Social Web
- Análise semântica de diferentes períodos
- Essa identificação dos temas que circundam a entidade analisada é feita através da análise de Tópicos ao Longo do Tempo (Topics-Over-Time, TOT) e K-means Clustering



Dataset: Google N-Gram

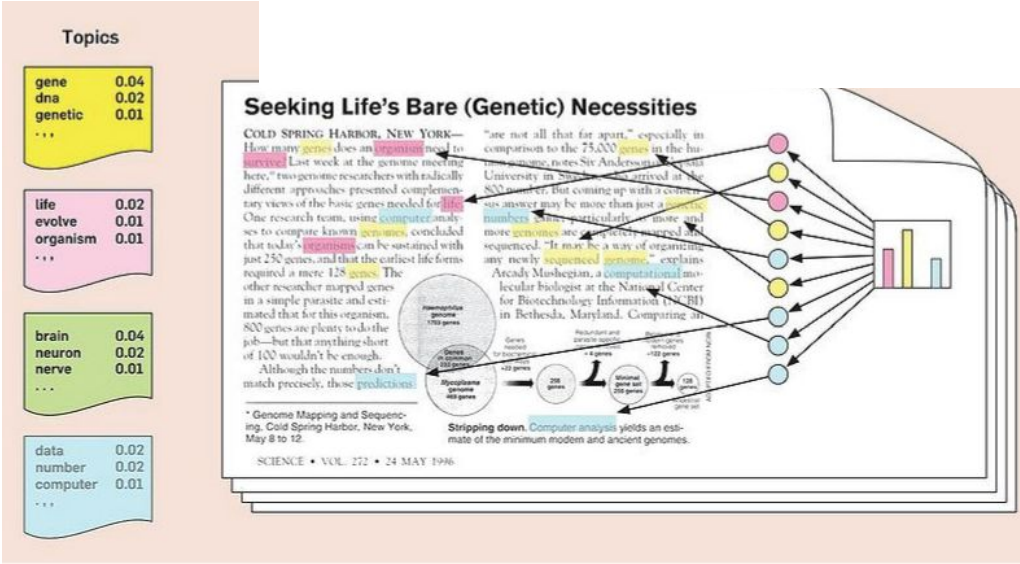
- 5-gramas em aprox. 800 arquivos de 1 GB.

- Nossa amostra:
    - Woman: 1,07 mi 5-gramas ~35MB
    - Man: 4,4 mi 5-gramas. ~130MB

Modelagem de tópicos

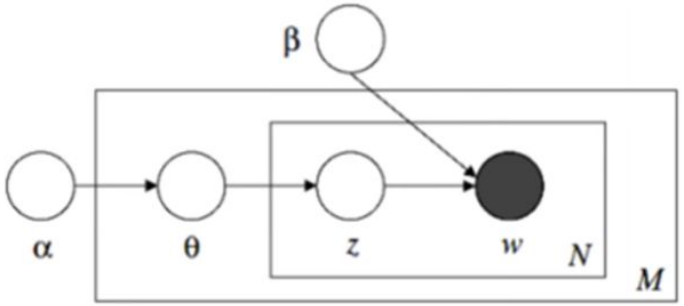
- PLN não supervisionada
- Documentos podem ser considerados uma coletânea de tópicos
- Análogo a clusterização, mas com combinações de palavras.

Detecta a temática oculta do texto



Latent Dirichlet Allocation

LDA

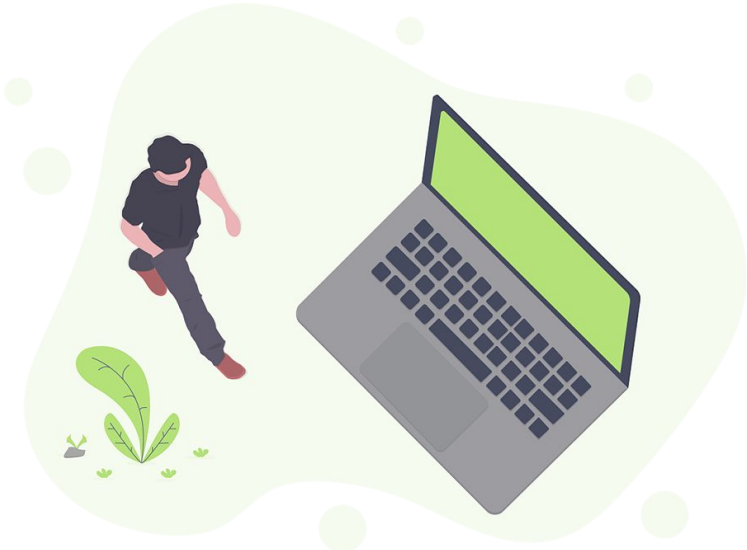


1. Aloca aleatoriamente um tópico para cada palavra
2. Para cada documento d:
   a. Assume que todos tópicos alocados estão corretos, exceto para d
   b. Calcula:
      i. Proporção de palavras no documento d que estão alocadas ao tópico t. *t = p(topico t | documeto d)*
      ii. Proporção de alocações ao tópico t em relação a todos documentos relacionados a palavra w. *w = p(palavra w | topico t)*
      iii. *t * w para alocar w a um novo tópico. Repete até chegar a um estado estável.*

Processamento: Para *woman* e *man*

- Seleciona as 2 palavras prévias e posteriores
- Criação de corpus para cada século
- Remoção caracteres [^A-Za-z]
- Transformação em minúsculas
- Tokenização
- Remoção Stopwords
- PorterStemmer
- Bag of Words
- Tf-idf
- LDA

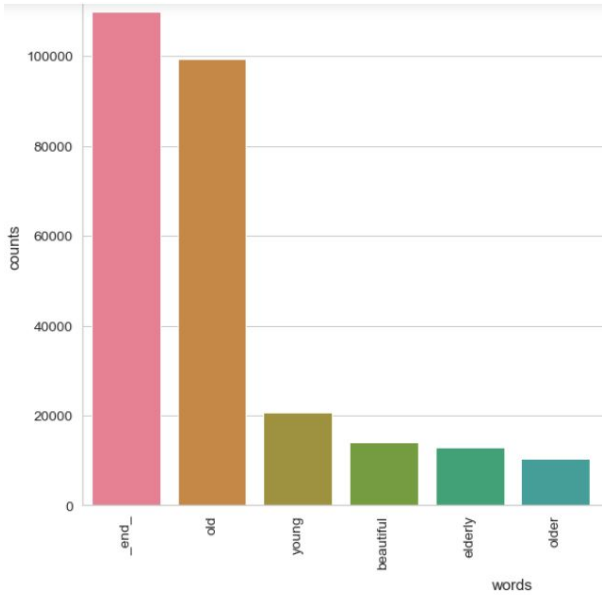

IF YOU MINE THE DATA HARD ENOUGH, YOU CAN ALSO FIND MESSAGES FROM GOD.
© 1999 United Feature Syndicate, Inc.
1/3/00



Palavras mais comuns

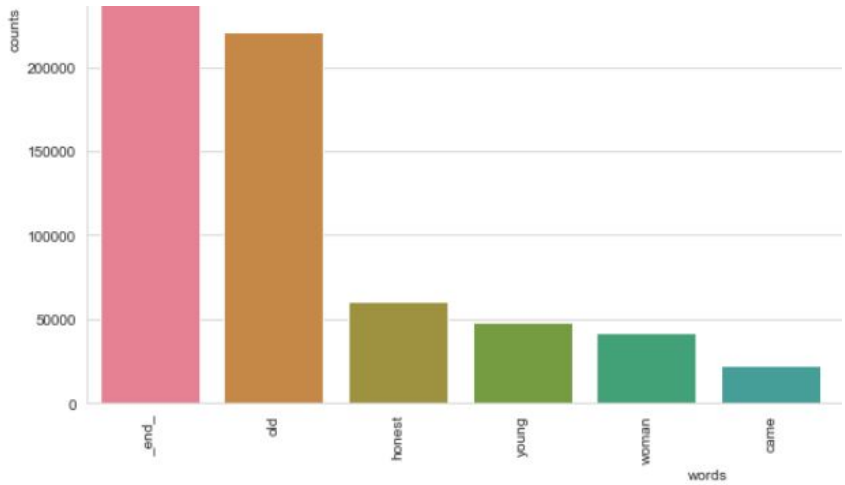
As palavras mais comuns não variam muito entre gênero no dataset

Woman



Old, Young, Beautifull, Elderly, Older

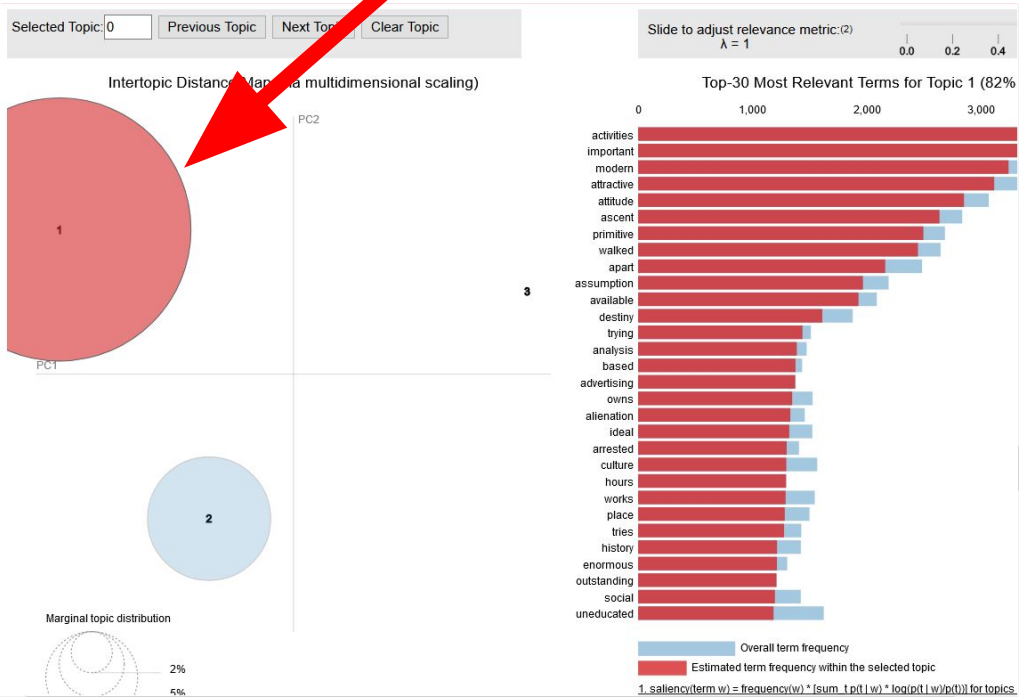
Man



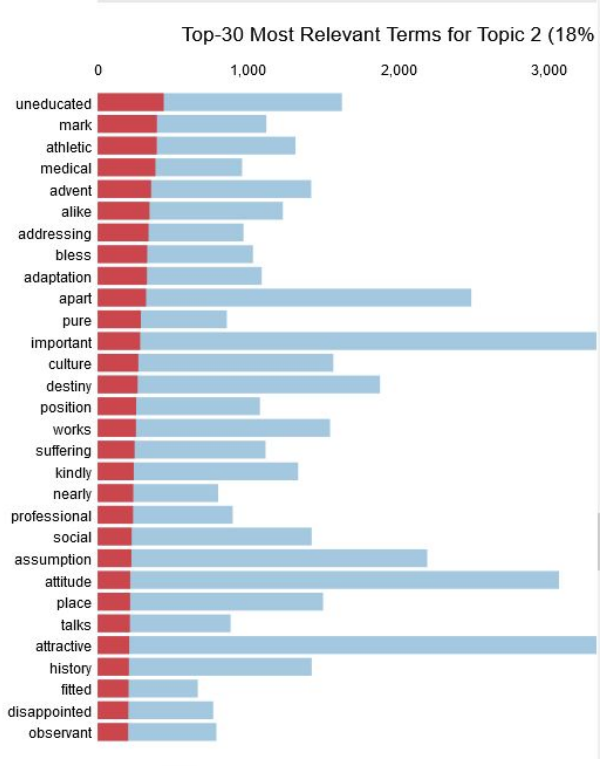
Old, Honest, Young, Woman, Came

Tópicos desbalanceados

Aqui tem 82% dos termos WTF?



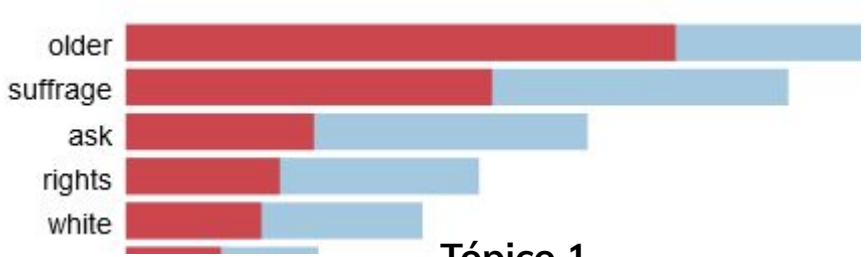
Variamos o número de tópicos de 3 a 7, porém os resultados continuaram desbalanceados



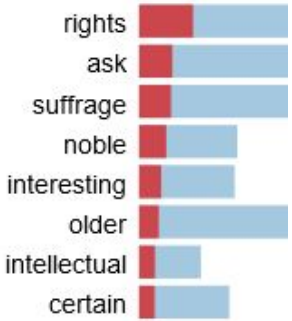
Porém verificamos variações interessantes. Homens estão mais relacionados a características externas e mulheres a questões internas/pessoais (ex: direitos)

Detecção do número de tópicos

Woman

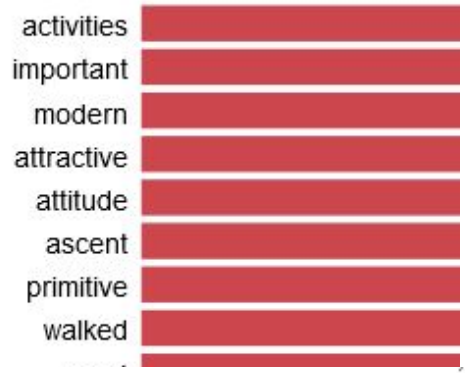


Tópico 1

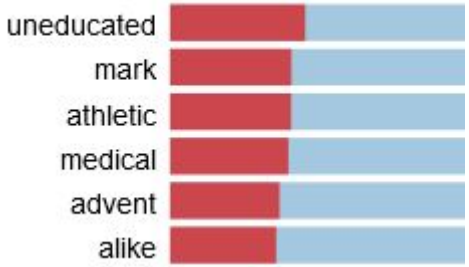


Tópico 2

Man



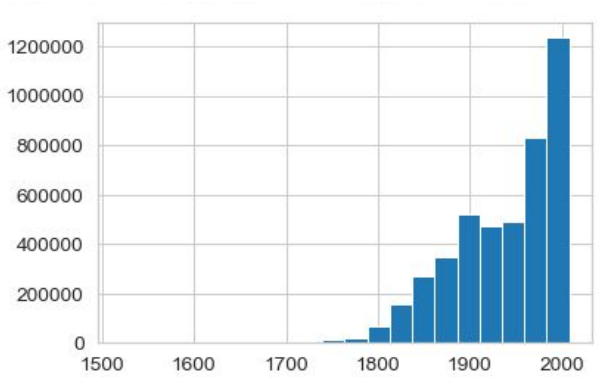
Tópico 1



Tópico 2

Curiosidades e limitações

O principal tópico relacionado a mulheres tem o termo sufrágio. Isso abre espaço para discutir o desbalanceamento dos dados no tempo.



Em uma das modelagens LDA para woman (n_topics = 3, anos=ALL, vectorizer=TF), **african** era o 3° termo mais importante do primeiro tópico (47,3%) e **white** o 5° principal do segundo (40,5%)

Problemas no TFIDF:

$$VolCount(t) \approx totalWords_t / aveWordsPerVolume_t$$
$$\approx \sum_x Match_t(x) / [(1/|x|) * \sum_x Match_t(x)/Vol_t(x)]$$

Variação de Tópicos ao longo dos séculos

VB = 9
ADJ = 17 (65%)

Woman

1500: 0.200*"honest" + 0.080*"love" + 0.080*"told" + 0.080*"infam" + ...

1600: 0.049*"born" + 0.037*"ask" + 0.037*"forget" + 0.037*"hath" + 0.037*"love" + 0.037*"honest"

1700: 0.028*"beauti" + 0.027*"young" + 0.025*"honest" + 0.021*"bond" + 0.020*"amiabl" + ...

1800: 0.028*"beauti" + 0.028*"young" + 0.013*"love" + 0.013*"born" + 0.013*"everi" + 0.013*"whose"

1900: 0.023*"young" + 0.019*"elderli" + 0.019*"ask" + 0.018*"older" + 0.016*"attract" + 0.016*"beauti"

2000: 0.023*"young" + 0.017*"anoth" + 0.017*"beauti" + 0.016*"ask" + 0.015*"elderli" + 0.015*"older"

Variação de Tópicos ao longo dos séculos

VB = 12
ADJ = 10 (45%)

Man

1500: 0.077*"honest" + 0.036*"great" + 0.036*"everi" + 0.036*"english" + 0.027*"might" + 0.019*"say"+ ...

1600: 0.060*"everi" + 0.054*"honest" + 0.025*"great" + 0.021*"may" + 0.019*"bodi" + 0.015*"good" + ...

1700: 0.067*"everi" + 0.043*"honest" + 0.017*"good" + 0.015*"may" + 0.013*"great" + 0.012*"anoth" + ...

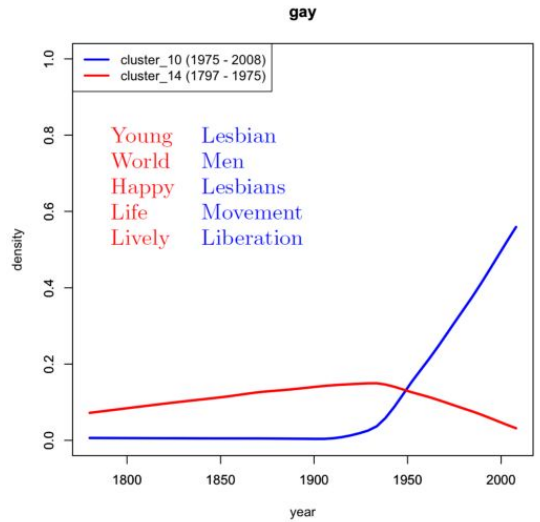
1800: 0.045*"everi" + 0.020*"honest" + 0.011*"call" + 0.011*"could" + 0.011*"anoth" + 0.010*"would" + ...

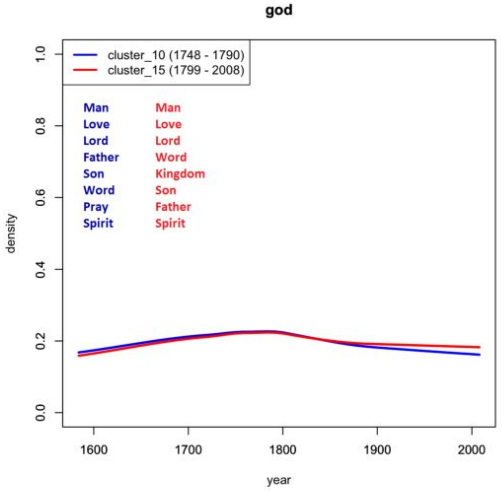
1900: 0.024*"everi" + 0.015*"anoth" + 0.013*"honest" + 0.012*"could" + 0.011*"ask" + 0.011*"woman" + ...

2000: 0.023*"everi" + 0.016*"anoth" + 0.012*"could" + 0.011*"call" + 0.011*"becom" + 0.011*"ask" + ...

Conclusões

- Enquanto que algumas palavras tem seu significado modificado ao longo dos séculos, outras mantém associadas a palavras similares.
- Assim como god, woman se encaixa nesse grupo de palavras estáticas. Apesar disso, detectamos algumas variações ao longo dos séculos, mas podem ser apenas resultante de sub-amostragem pois isso ocorreu comparando séculos 15, 16 e 20.
- Palavras associadas a homem e mulher, similares quanto a beleza e juventude.
- Honestidade aparece mais associada a homem.
- O trabalho original reporta apenas clusters selecionados manualmente em comparações pareadas, o que pode indicar algum viés de seleção.





Bibliotecas utilizadas

Google_ngram_downloader - Dataset

Nltk - tokenização, stopwords, PorterStemmer

Re

Seaborn e pyLDAvis - Visualização

Sklearn - LDA, CountVectorizer, TfidfVectorizer

Gensim - LDA, tf-idf

Vocab, man = 9234 / woman = 3637

Google Acadêmico

topic over time

Artigos

Aproximadamente 1.410.000 resultados (**0,08** s)

Meu perfil

A qualquer momento
Desde 2019
Desde 2018
Desde 2015
Período específico...

— 2010

Pesquisar

Classificar por relevância
Classificar por data

Em qualquer idioma
Pesquisar páginas em Português

incluir patentes
incluir citações

Criar alerta

Topics **over time**: a non-Markov continuous-**time** model of topical trends

[PDF] academia.edu

X Wang, A McCallum - Proceedings of the 12th ACM SIGKDD …, 2006 - dl.acm.org

… Second, many other models take the view that the “mean- ing” (or word associations) of a **topic** changes **over time**; in- stead, in TOT we can rely on topics themselves as constant, while **topic** co-occurrence patterns change **over time** …

☆ Citado por 1321 Artigos relacionados Todas as 18 versões

Dynamic **topic** models

[PDF] psu.edu

DM Blei, JD Lafferty - Proceedings of the 23rd international conference …, 2006 - dl.acm.org

… A A A θ θ θ z z z α α α β β β w w w N N N K Figure 1. Graphical representation of a dynamic **topic** model (for three **time** slices). Each **topic's** natural parameters βt,k evolve **over time**, together with the mean parameters αt of the logistic normal distribution for the **topic** proportions …

☆ Citado por 2180 Artigos relacionados Todas as 18 versões

[PDF] An automated method of **topic**-coding legislative speech **over time** with application to the 105th-108th US Senate

[PDF] researchgate.net

KM Quinn, BL Monroe, M Colaresi… - Midwest Political …, 2006 - researchgate.net

In this paper, we describe a method for statistical learning from speech documents that we apply to the Congressional Record in order to gain new insight into the dynamics of the political agenda. Prior efforts to evaluate the attention of elected representatives across **topic** …

☆ Citado por 70 Artigos relacionados Todas as 4 versões »

Topic and trend detection in text collections using latent dirichlet allocation

[PDF] psu.edu

L Bolelli, Ş Ertekin, CL Giles - European Conference on Information …, 2009 - Springer

… segments. Our experimental results on a collection of academic papers from CiteSeer repository show that segmented **topic** model can effectively detect distinct topics and their evolution **over time**. 1 Introduction and Related Work …