# *Text Mining- Modelação de Tópicos*

Trabalho 2 – 2º Semestre (2017/2018)

Professores : Ricardo Ribeiro e Fernado Batista

Mariana Rebelo Dias – 68494
Mestrado em Engenharia Informática- SIGC
ISCTE, Lisboa
mrdso@iscte-iul.pt

Marco Filipe Madeira Felgueiras - 54382
Mestrado em Engenharia Informática- SIGC
ISCTE, Lisboa
mfmfs@iscte-iul.pt

***Abstract*—A Modelação de Tópicos consiste é uma tarefa de Classificação de Texto, com o objetivo de obter os tópicos relativamente a um texto, ou seja classificar um texto de acordo com a sua categoria/ assunto. Este trabalho tem como objetivo fomentar a compreensão do que é um modelo de tópicos e a extracção a partir do conjunto de dados The SFU_Review_Corpus, isto através de diferentes experiências de criação de modelos de tópicos, como LSA, LDA e HDP, aplicação dos mesmos e avaliação realizando inferência. Assim como a análise da influência da forma de representação dos documentos em tarefas de criação de tópicos, fazendo diversas experiencias ao nível do pré-processamento.**

**Keywords—Topic Modeling, Latent Semantic Analysis, Latent Dirichlet Allocation, Hierarchical Dirichlet process, Topic Detection.**

## I. INTRODUÇÃO

A deteção de tópicos é uma tarefa da área do NLP, que consiste em atribuir um tópico a um conjunto de termos de um determinado documento [6], é, por norma, constituída por algoritmos de aprendizagem não supervisionada que utilizam técnicas de *clustering* para encontrar *latent variables* ou estruturas ocultas dentro dos dados. O trabalho proposto decompõe-se em 4 fases, a preparação dos dados, a construção dos modelos de tópicos, a representação dos documentos e a inferência.

## II. ANÁLISE DO TRABALHO RELACIONADO

A grande maioria do trabalho realizado na área de *Opinion Mining* e *Topic Modeling* concentra-se em tarefas de identificação e baseia-se em *features* [1], estando este tema em constante crescimento maioritariamente nos últimos anos.

A primeira abordagem consistiu em procurar modelos de tópicos baseados em *bag-of-words*. Os autores em [2] mostram diferentes modelos de tópicos todos baseados no modelo probabilístico LDA (Latent Dirichlet Allocation), para fazer modelação de tópicos e um estudo focado em opiniões.

Por outro lado em [3] a abordagem tomada relaciona a modelação de tópicos focada em aprendizagem de *Latent Aspects* de forma a utilizar um conjunto de palavras para descrever um determinado aspeto/palavra. Esta descoberta de palavras é combinada ao modelo LDA.

Outros autores utilizam além do LDA para modelação de tópicos também o modelo LSA (Latent Semantic Analysis) [4], considerando-os dois modelos standard para este tipo de tarefas. Assim como também HDP (Hierarchical Dirichlet process) combinado com *latent topic* e n-grams [5]

Ao nível do da representação de documentos é muito utilizado a remoção de sinais de pontuação e de palavras sem contexto [6] assim como tentar reduzir ao máximo palavras irrelevantes à descoberta de tópicos como o exemplo as stop words.

## III. DESCRIÇÃO DOS DADOS E MÉTRICAS

Para a realização deste trabalho foi utilizado o conjunto de dados The SFU Review Corpus. Este conjunto de dados é composto por diferentes reviews em Inglês retiradas do site Epinions em 2004.

No total o conjunto de dados utilizado é composto por 400 opiniões de utilizadores relativamente a 8 categorias distintas: Livros, Carros, Computadores, Utensílios de Cozinha, Hotéis, Filmes, Musicas e Telemóveis. Sendo o conjunto de dados bem divido e composto por um total de 50 reviews de cada uma das categorias.

Os dados consistem num documento *json* composto por dois parâmetro: “text” e “recommended”. O primeiro parâmetro “text” diz respeito ao texto da review que será analisado, o segundo é ignorado neste caso uma vez que não é relevante para este tipo de análise. Podemos ver que não existe no conjunto de dados original uma classificação relativamente aos tópicos que possibilite a comparação com os resultados dos modelos de tópicos ou no futuro uma aplicação de métodos de aprendizagem supervisionada. De forma a solucionar esse problema realizamos manualmente a categorização de cada um dos documentos e construímos um novo conjunto de dados composto agora pelos parâmetros “text” e “topic”. Em que o “text” mantém-se sem alterações relativamente ao anterior e o novo parâmetro “topic” corresponde ao tópico associado a cada uma das *reviews*. Assim conseguimos um conjunto de dados de diferentes reviews e o tópico correspondente a cada uma delas.

Além da inferência e análise dos tópicos optamos também por utilizar uma métrica de classificação de tópicos. Como métrica de avaliação de resultados é utilizada a coerência. A coerência de um tópico mede a “interpretabilidade” humana de um conjunto de tópicos. Esta métrica baseia-se na segmentação, cálculo de probabilidades considerando o corpus utilizado, numa medida de confirmação e agregação de termos. A coerência como uma medida de avaliação/comparação de

diferentes modelos de tópicos oferece garantia muito superior de "interpretabilidade" humana comparando com outras medidas que correlacionam tópicos.

## IV. TRABALHO PROPOSTO

A descrição das tarefas desenvolvidas ao longo deste trabalho assim como a descrição de todas as opções tomadas, divide-se neste ponto nas quarto tarefas do trabalho: Preparação dos Dado, Construção dos modelos de tópicos, Representação dos documentos e Inferencia.

### *A. Preparação dos Dados*

Esta primeira tarefa consiste numa análise e preparação simples relativamente ao conjunto de dados.

Uma vez que os dados ficam organizados por categoria é necessário começar por desordenar todo o conjunto de forma a ter um conjunto aleatório de *reviews* das diferentes categorias. Foi também necessário fazer uma divisão dos dados de forma a guardar 10 documentos para posteriormente se realizar inferência de tópicos.

Com isto o conjunto de dados sobre o qual se vai trabalhar é composto por um total de 390 *reviews* com a seguinte divisão de categorias:

TABLE I. DIVISÃO DO CONJUNTO DE DADOS

| Categoria | Número de Reviews |
|---|---|
| Books | 48 |
| Cars | 49 |
| Computers | 49 |
| Cookware | 49 |
| Hotels | 48 |
| Movies | 50 |
| Music | 48 |
| Phones | 49 |

Ao nível da preparação dos dados são realizadas apenas tarefas simples como eliminar a pontuação e caracteres especiais, assim como também são retirados os *linebreakers* ("\n","\r") presentes em todos nos documentos. Além disso e uma vez que são palavras sem significado para a criação de tópicos são eliminadas também todas as *stopwords*.

Finalmente é construído um dicionário de termos/palavras e uma matriz de documentos por termos.

### *B. Construção dos modelos de tópicos*

Esta tarefa visa a criação de modelos de tópicos e uma *word cloud* utilizando o conjunto de dados reservado para treino. Na construção dos modelos de tópicos os dados encontram-se apenas com o pré-processamento referido no ponto acima.

O primeiro modelo de tópicos é construindo utilizando o LSA (Latent Semantic Analysis) da biblioteca gensim, o único parâmetro dado ao modelo é: o número de tópicos no qual utilizamos o valor 8.

O outro modelo de tópicos diz respeito ao LDA (Latent Dirichlet Allocation) também construído utilizando a biblioteca do gensim. Neste modelo além do parâmetro número de tópicos ao qual atribuímos os valor 8, tal como no anterior. Existem mais dois parâmetros que foram necessários estudar de forma a perceber como a sua variação influencia os tópicos produzidos pelo modelo. Foi utilizado para o número de passagem o valor 20 e para o número de iterações, outro parâmetro do modelo, um número de iterações igual a 100.

Após a criação dos dois modelos é possível passar à criação de uma *word cloud* para cada um dos modelos de tópicos. A *word cloud* é construída de acordo com as palavras obtidas em cada modelo e a informação sobre a sua importância. Através das *word clouds* conseguimos verificar quais as palavras mais importantes e se o modelo consegue produzir informação que permita a categorização de tópicos para realizar inferência.

Por fim é feita uma análise aos modelos anteriormente criados de forma a perceber se é possível atribuir alguma categorização a cada conjunto de palavras. E como última experiência na tentativa de perceber qual a melhor forma de abordar o problema e estudar novos resultados, variámos o número de tópicos para os dois modelos para metade e para o dobro, ou seja, são criados novamente os modelos LSA e LDA mas com o número de tópicos 4 e 8.

### *C. Representação dos documentos*

Por forma a tentar melhorar os resultados obtidos no ponto anterior nesta fase são feitas várias experiencias ao nível da forma de representação de documentos e da construção de diferentes modelos de tópicos.

Nesta fase, além do pré-processamento realizado anteriormente em que só era retirada a pontuação, *linebreaks* e eliminadas as *stop-words*, e uma vez que apenas este tratamento não nos dá informação suficiente relativamente aos tópicos optámos por diferentes abordagens na forma de representação dos documentos. Fizemos experiências como:

- retirar os números, uma vez que estes não nos dão informação relativa a tópicos e eram bastante frequentes nos mesmos;
- Part-Of-Speech tagging, considerando apenas nomes;
- Stemming;
- Lematização;
- duas abordagens de *Chunking*, a primeira em que apenas são considerados os chunks dos documentos na criação dos modelos de tópicos, como por exemplo na frase: "Compared to the work of more competent practitioners" ficaríamos só com ['work', 'competent practitioners']. Na segunda abordagem é considerado todo o texto das *reviews* mas substituíndo os chunks respectivos, por exemplo na mesma frase do exemplo anterior neste caso ficaríamos com: ['Compared ', 'to', 'the','work', 'of', 'more', 'competent_practitioners'].
- TF-IDF, onde apenas consideramos as 6000 palavras com maior IDF e ao mesmo tempo cujo TF seja menor ou igual que 195 ocorrências, ou seja metade do número de documentos de treino. Isto de forma a tentar reduzir palavras que estão presentes em vários documentos e não são características de um tópico e considerar apenas as mais importantes que nos possam dar informação relevante na descoberta dos tópicos.

Algumas destas experiências são aplicadas em simultâneo aos documentos de forma a tentar ter um conjunto de palavras em cada tópico mais adequado e coerente.

Além dos dois modelos construídos, LSA e LDA optamos pela construção de um novo modelo de tópicos, HDP. Todas as experiências nesta fase são realizadas com os 3 modelos de

tópicos e com apenas o conjunto de dados reservado para treino, ou seja 390 documentos.

### *D. Inferência*

Para aplicar os modelos de tópicos aos 10 documentos reservados para realizar inferência realizamos o mesmo pré-processamento que foi aplicado aos documentos de treino, convertendo depois todos os termos que os compõem em identificadores numéricos para que estes possam ser alimentados aos diferentes algoritmos (LSA, LDA, HDP). Todos os algoritmos referidos emitem para um dado documento um valor associado a cada tópico. Para este valor consideramos o valor absoluto por forma a tentar encontrar qual o maior "valor de confiança" associado a cada tópico, posteriormente foi feita a comparação com a categoria real de cada *review*.

## V. IMPLEMENTAÇÃO E RESULTADOS

Conforme descrito, todos os dados passam por uma preparação inicial onde é realizado um pré-processamento simples, a divisão dos documentos de treino e teste e construída uma matriz de documentos por termos. Foram construídos os dois modelos pedidos, LSA e LDA com o conjunto de dados reservados para treino onde obtivemos os seguintes valores de coerência:

TABLE II. RESULTADOS DOS MODELOS DE TÓPICOS

| Modelo | Valor da Coerência |
|---|---|
| LSA | 0.38563243077915904 |
| LDA | 0.300336580141537 |

Conseguimos ver pelos valores de coerência que o LSA obteve resultados melhores que o LDA, mas ainda assim pela avaliação dos tópicos verificamos que em ambos é bastante difícil inferir alguma das categorias.

TABLE III. CORRESPONDENCIA DE TÓPICOS LSA

| Tópico | Categorias |
|---|---|
| [ like , one , car , would , also , get , good , much , time , iMac , even , really , dont , well , new , Dell , use , back , want , two ] | Cars/ Computers |
| [ car , iMac , Mac , Dell , PC , Apple , computer , photos , iPhoto , applications , rear , seat , page , 1 , engine , cars , machine , seats , mouse , easy ] | Cars/ Computers |
| [ car , track , song , beat , Ras , album , lyrics , Kass , Stars , rap , chorus , Dell , system , spits , one , us , hip , 5 , hop , production ] | Music |
| [ Dell , 1 , system , iMac , Mac , Customer , 2 , may , Care , software , System , Apple , photos , drive , iPhoto , problems , performance , computer , Ras , applications ] | Computers, Phones |
| [ car , Ras , AllClad , iMac , track , Kass , Mac , room , phone , one , pan , Stainless , set , hotel , PC , book , Apple , album , dont , photos ] | Computers, Cookware, Hotels |
| [ Ras , Kass , song , beat , spits , Murphy , Ice , chorus , Soul , us , hop , hip , lyrics , Lee , Nelly , album , room , verse , metaphors , Lil ] | Music |
| [ AllClad , Stainless , pan , Steel , Fry , cookware , Pan , room , pans , stainless , book , set , use , heat , Pans , hotel , movie , film , kitchen , steel ] | Cookware, Hotels |
| [ phone , room , handset , Panasonic , phones , battery , cordless , base , hotel , handsets , Disney , system , features , unit , resort , ID , use , caller , station , pool ] | Hotels/ Phones |

TABLE IV. CORRESPONDENCIA DE TÓPICOS LDA

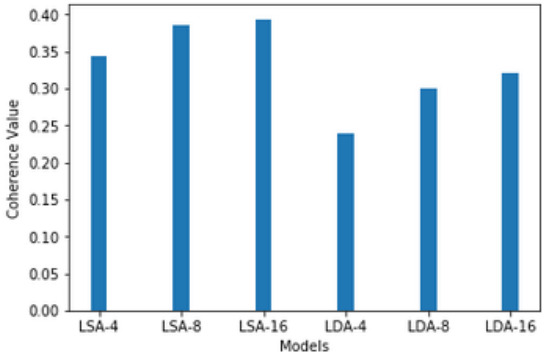
| Tópico | Categorias |
|---|---|
| [ room , hotel , Disney , resort , pool , stay , rooms , time , two , one , also , Club , get , like , Beach , good , night , area , take , go ] | Hotels |
| [ song , album , one , track , like , beat , Dell , lyrics , good , 1 , 5 , rap , songs , Ras , music , even , get , great , time , Stars ] | Music |
| [ one , computer , would , like , book , get , use , also , dont , time , iMac , Apple , even , pan , much , AllClad , new , pans , little , set ] | Computers |
| [ phone , handset , one , phones , like , use , Panasonic , good , base , would , system , battery , dont , cordless , set , 2 , handsets , features , time , back ] | Phones |
| [ like , one , movie , get , Samurai , even , book , good , would , go , time , really , well , want , though , see , way , dont , Cruise , say ] | Movies/ Books |
| [ movie , like , film , one , track , album , song , would , beat , make , kids , first , Murphy , think , lyrics , way , Buddy , comes , even , plot ] | Movies |
| [ like , book , one , Stephanie , movie , much , way , dont , story , never , read , really , know , good , would , time , get , well , could , books ] | Books |
| [ car , like , one , engine , also , would , get , good , cars , rear , power , even , seat , much , seats , drive , front , driving , Ford , 2002 ] | Cars |

Nas tabelas III e IV podemos ver que existem ainda algumas palavras que não são muito informativas e que não levam a uma fácil compreensão dos tópicos, como por exemplo os números ou os pronomes. Conseguimos ver também que apesar dos valores de coerência serem mais altos no LSA ambos são por vezes bastante difíceis de categorizar ou de distinguir entre tópicos.

Após a criação dos dois modelos passamos a criação das *word clouds*, como forma de representação da importância das palavras no universo de tópicos. A geração das mesmas foi feita a partir da biblioteca *WordCloud* do *python* passando-lhe uma estrutura com a palavra e respectivo valor associado dentro do conjunto de tópicos. As duas *word clouds* construídas, tanto para o LSA como para o LDA podem ser vistas em anexo neste documento.

Para a variação do número de tópicos nos dois modelos optamos como já referido por utilizar metade e o dobro do número de tópicos existentes. Os valores de coerência obtidos podem ser vistos no Gráfico I a baixo:

GRAPH I. RESULTADOS DA VARIAÇÃO DO NUMERO DE TOPICOS




Após a análise do gráfico correspondente aos valores de coerência e às palavras geradas pelos modelos de tópicos,

podemos ver que apesar de os melhores resultados serem para 16 tópicos em ambos os modelos a diferença não é significativa ao nível da coerência e quando tentamos analisar as palavras geradas pelos tópicos tornou-se ainda mais difícil de inferir um tópico quando variado o número de tópicos para 16. O mesmo acontece quando fazemos a redução de tópicos para 4 tópicos, mas com a diferença que os resultados a nível de coerência são um pouco inferiores.

Podemos concluir desta experiencia que a tentativa de aumento e redução do número de tópicos não ajudou à criação de melhores tópicos para inferência.

Como tentativa de melhorar os resultados dos diferentes modelos de tópicos foram aplicadas várias experiencias de representação de documentos, tais como:

-Part-Of-Speech Tagging, de todas as frases dos documentos foram extraidos apenas os termos cuja tag fosse NN, NNP, NNPS ou NNS;

-Stemming, após a extracção dos termos por parte do POS, todos estes foram reduzidos ao seu stem utilizando o PorterStemmer da biblioteca NLTK;

-Lematização, aplicado tal como o *stemming*, após a extracção dos termos por parte do POS, todos estes foram reduzidos ao seu *lemma* utilizando o WordNetLemmatizer da biblioteca NLTK;

-Chunking:

-Primeira abordagem, através da biblioteca spaCy, todas as reviews foram analisadas no seu motor de NLP, utilizando o módulo *en_core_web_sm*, a partir daqui o *spaCy* é capaz de extrair um conjunto alargado de informação de um determinado texto. Para representação final de cada documento, foram extraídos todos os seus *noun_chunks*;

-Segunda abordagem, semelhante à abordagem anterior, contudo, apenas foram substituídos os *chunks* nos documentos, como explicado na secção anterior;

-TF-IDF: para a implementação do TF-IDF foi utilizado o *TfidfModel* da biblioteca *gensim*, este serviu para criar um vocabulário a partir dos documentos de treino, posteriormente todos os termos de todos os documentos só são considerados se estiverem neste mesmo vocabulário; Além disto, optamos ainda por combinar todas as abordagens acima descritas com a implementação do TF-IDF.

Adicionalmente, para todas estas experiências foi ainda aplicado o HdpModel da biblioteca gensim.

Os resultados obtidos pela aplicação das diferentes experiências de representação de documentos e dos 3 modelos aplicados podem ser vistos na tabela abaixo:

TABLE V. RESULTADOS DAS EXPERIENCIAS REALIZADAS

| | **LSA** | **LDA** | **HDP** |
|---|---|---|---|
| POS | 0,404 | 0,366 | 0,366 |
| POS+Stemm | 0,425 | 0,326 | 0,357 |
| POS+Lem | 0,416 | 0,407 | 0,349 |
| Chunking -1 | **0,448** | 0,365 | 0,451 |
| Chunking -2 | 0,444 | 0,238 | 0,278 |
| TF-IDF | | | |
| POS | 0,405 | 0,717 | **0,805** |
| POS+Stemm | 0,405 | 0,729 | 0,794 |
| POS+Lem | 0,399 | 0,709 | 0,788 |
| Chunking -1 | 0,416 | **0,731** | **0,804** |

Analisando os valores obtidos ao nível da para cada um dos modelos de tópicos conseguimos perceber que os melhores resultados são para a abordagem que utiliza chunking com TF-IDF nos modelos LDA e HDP. Tendo tido estes dois modelos uma grande variação nos resultados quando aplicado o TF-IDF. O mesmo não ocorre ao nível do LSA que independentemente do tipo de pré-processamento dos dados nunca se altera muito. Apesar dos resultados da coerência terem algumas alterações significativas o mesmo não se reflecte nos termos gerados para cada tópico, sendo na maioria da vezes bastante difícil de inferir um tópico dos conjuntos de palavras.

Após realizada esta experiencia e uma vez que o resultado mais elevado foi utilizando a primeira abordagem do *chunking* com TF-IDF e sendo também onde é mais perceptível qual o tópico em cada conjunto de termos, optamos por realizar inferência para os 10 documentos previamente reservados para o efeito através desta abordagem.

Na tabela abaixo podemos ver o resultado obtido pela realização de inferência a todos os modelos:

TABLE VI. RESULTADOS DA INFERENCIA DE TOPICOS

| **Doc** | **LSA** | **LDA** | **HDP** | **Categ. original** |
|---|---|---|---|---|
| 0 | 7-Cars | 5-Music | 47-?? | Books |
| 1 | **1-Music** | 7- Cookware/Hotels | **53-Music** | Music |
| 2 | 7- Cars | 6- Music/Movies | **31-Hotel** | Hotel |
| 3 | 4-Computers | 3-?? | 15-?? | Cars |
| 4 | **5- Music** | **6- Music/Movies** | 29-?? | Music |
| 5 | **2-Cookware** | **7-Cookware/Hotels** | 93-?? | Cookware |
| 6 | **4- Computers** | **2- Computers** | 77-Cars | Computers |
| 7 | 0-Music | 4-Music | 69-Phones | Hotels |
| 8 | 6- Computers | **0-Books** | **88-Books** | Books |
| 9 | 6- Computers | 4- Music | 114-?? | Phones |

Através da inferência pelos 10 documentos vimos que ambos os modelos se comportaram da mesma forma, tendo poucas diferenças entre eles. Ainda assim quando comparado o tópico atribuído pelos modelos com a categorização original a percentagem de acerto foi bastante baixa.

## VI. CONCLUSÕES E PROPOSTAS FUTURAS

Foram realizadas diversas experiencias a nível do pré-processamento do conjunto de dados e utilizados vários modelos de tópicos, contendo várias variações dentro dos mesmos e tentando diferentes abordagens. Contudo, com todas as tentativas de melhoria de resultados efectuadas, o ganho não foi muito significativo. É possível que isto se deva ao conjunto de dados que estamos a utilizar, por este motivo, futuramente seria necessário aplicar os mesmos algoritmos a um conjunto de dados diferente, para verificar o comportamento dos mesmos.

Gostaríamos também de ter tido mais tempo para testar outras abordagens como por exemplo o uso dos embeddings e o tratamento de sinónimos uma vez que acreditamos que podem influenciar positivamente os resultados.

## REFERENCES


[1] u, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177)

[2] Moghaddam, S., & Ester, M. (2012, October). On the design of LDA models for aspect-based opinion mining. In Proceedings of the 21st ACM international conference on Information and knowledge management (pp. 803-812)

[3] Moghaddam, S., & Ester, M. (2011, July). ILDA: interdependent LDA model for learning latent aspects and their ratings from online product reviews. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval (pp. 665-674)

[4] Titov, I., & McDonald, R. (2008, April). Modeling online reviews with multi-grain topic models. In Proceedings of the 17th international conference on World Wide Web (pp. 111-120)

[5] Wallach, H. M. (2006, June). Topic modeling: beyond bag-of-words. In Proceedings of the 23rd international conference on Machine learning (pp. 977-984)

[6] Batista, F., & Ribeiro, R. (2013). Sentiment analysis and topic classification based on binary maximum entropy classifiers. Procesamiento del lenguaje natural, (50), 77-84

Anexos

Word cloud do modelo LSA


Word cloud do modelo LSA