Topic Modeling Reviews
Text Mining 2019-2020

Tiago Almeida, 69362
MEI-SIGCA1
ISCTE-IUL
tfpaa@iscte-iu.pt

Patricia Santos, 64732
MEI-SIGCA1
ISCTE-IUL
psssa@iscte-iul.pt

***Abstract*— Neste artigo vamos abordar o tema de Topic Modeling em reviews de múltiplas categorias, tais como, hotéis, musica, livros, carros, computadores, entre outras. Relativamente ao tratamento de dados, implementámos uma junção de dois tratamentos, um direcionado ao texto da review e outro aplicando técnicas de processamento da língua, como por exemplo, Lemmatization, POS Tagging e Chunking. Depois desta etapa, treinamos os modelos LDA, LSA, criados por nós e o HDP. Por fim, numa tentativa de melhorar os resultados, vamos experimentar diferentes tipos de pré-processamento. Como extra nesta análise, vamos estudar a relação do número de tópicos com a coherence, de maneira a saber com quantos tópicos obtemos o valor máximo da coherence.**

Keywords*— *topic modeling, reviews, lda, lsa, Latent Dirichlet Allocation, Latent Semantic Analysis, hdp

I. INTRODUÇÃO

Hoje em dia, quando uma pessoa está perante uma indecisão nas escolhas que quer fazer, tem por hábito basear-se e influenciar-se nas recomendações, classificações e *reviews* dadas por outros utilizadores, especialmente se for em assuntos ou situações das quais tem pouco ou nenhum conhecimento, como por exemplo, se um certo filme é bom, se a relação qualidade/preço de um restaurante é justa, se o serviço e estadia num hotel foi agradável, entre outros. [1]

Contudo, devido à quantidade excessiva de opiniões é difícil separar quais aquelas que são realmente relevantes e importantes para o que o utilizador procura. Para este problema precisamos de algo que nos ajude a organizar e resumir toda a informação encontrada, ou seja, ao usar *Topic Modeling*, uma ferramenta do *Natural Language Processing* (mais conhecido por *NLP*), podemos encontra um grupo de tópicos que melhor representam a informação contida nos documentos analisados, extraindo as palavras mais recorrentes ou as que têm maior importância para determinado tópico.

No caso do nosso artigo, vamos aplicar esta ferramenta a um conjunto de *reviews* e fornecer aos utilizadores a informação mais relevante, para que se seja possível ajudar a tomar determinada decisão.

II. ANÁLISE DE LITERATURA

Neste capítulo vamos analisar artigos publicados sobre modelos de tópicos.

Relativamente ao modelo *Latent Dirichlet Allocation* (será referido como *LDA*) os autores do artigo [2] têm como objetivo criar um modelo generativo para texto, para isso utilizam técnicas como *unigrams, n-grams, Probabilistic Latent Semantic Indexing* (também conhecido por *pLSI*) e *LDA*. Afirmam que neste modelo se assume que todos os tópicos geram documentos, sendo que cada tópico tem uma representação semelhante a uma distribuição multinominal sobre as palavras e que, gerar um documento provêm da junção dos tópicos e palavras.

Outro artigo que analisa o modelo *LDA*, foi o dos autores [3] que consideram as palavras, *n-grams* e pesos *tf-IDF* as características mais comuns para classificação de texto, contudo, afirmam que as mesmas contêm problemas e, por isso, propuseram-se a criar uma representação de um *Vector Space Model* (daqui para a frente será mencionado como VSM) com *LDA* e espaços semânticos. Em primeiro lugar, aplicaram *Tokenization* e *Lemmatization* para calcular a frequência das palavras, de seguida, obtiveram os *n-grams* e determinaram os valores de *tf-IDF* e, por fim criaram um modelo *LDA* com diferentes números de tópicos e 5 VSMs, unigrams com e sem *stop words*, *bigrams* com e sem *stop words* e por último *Syntatic bigrams*. Concluíram que o modelo *LDA* tem uma maior taxa de sucesso só com um tópico por documento e quando é usado juntamente com VSM ao invés de ser usado individualmente.

Por fim, os autores do artigo [4] utilizaram uma abordagem diferente, combinando o modelo *LDA* com o *Topic-in-Set Knowledge* e, como conclusão demonstraram que o modelo *LDA* juntamente com o *z-labels* do *Topic-in-Set Knowledge* recupera tópicos mais interessantes.

Quanto ao modelo *Latent Semantic Analysis* (será referido como LSA) começamos pelo artigo [1] onde se propõe uma abordagem baseada neste modelo com o objetivo de reduzir o tamanho do resumo dos documentos, para identificar tópicos relacionados com produtos e recursos. Os autores do artigo caracterizam o modelo LSA como uma teoria e método que ajuda a analisar a relação de documentos com termos que eles possuem, gerando um conjunto de tópicos relacionados aos documentos e termos. Para conseguir alcançar o proposto, usaram um *dataset* de *reviews* de filmes chineses do *Internet Blogs* como *input* de um classificador de sentimentos, mais especificamente, SVM que classifica as *reviews* como positivas ou negativas. Após as experiências os autores concluíram que o modelo LSA pode identificar um conjunto de tópicos relacionados com o produto.

Por fim, o artigo [5] pretende apresentar uma estrutura para extrair aspetos quantificáveis a partir de *reviews* de utilizadores *online* utilizando modelos *multi-grain*, ao invés dos métodos de modelação de tópicos *standards* presente nos outros artigos, uma vez que, na visão dos mesmos os modelos standard tendem a produzir tópicos que correspondem às características gerais dos objetos em vez dos atributos do objeto classificado pelo utilizador. Depois das experiências aplicando modelos PLSA e LDA em diferentes *datasets* como por exemplo, *reviews* de música, *reviews* de hotéis e *reviews*

de restaurantes, concluíram que o modelo *multi-grain* são superiores para extrair tópicos mais precisos de *reviews online*.

III. PREPARAÇÃO DOS DADOS

O conjunto de dados escolhido para análise foi o "SFU_Review_Corpus.json". Este *dataset* é constituído por 400 *reviews* obtidas em 2004 do *site Epinions*. As *reviews* presentes no *dataset* dizem respeito a várias categorias como por exemplo, *"Books, Cars, Computers, Cookware, Hotels, Movies, Music, Phones"*.

O *dataset* encontra-se no formato *JSON* onde, cada linha é constituída por dois parâmetros, "*recommended*" que irá ser excluído da nossa análise, visto não ser relevante para o nosso objetivo e, "*text"* que irá conter a *review* sobre uma das categorias mencionadas anteriormente e ser o foco da nossa análise. Das 400 *reviews* que temos disponíveis no nosso *dataset* vamos dividir, 390 para treino dos modelos criados e 10 para inferência.

Para ajudar à nossa análise e identificação das categorias por tópicos, decidimos criar 8 ficheiros/bibliotecas com as palavras mais relacionadas com cada categoria acima mencionada. O preenchimento destas bibliotecas foi com base no *site RelatedWord* [1], que utiliza vários algoritmos para obter as palavras mais frequentes para a palavra pesquisada, que no nosso caso esta palavra corresponde ao nome da categoria. Dois desses algoritmos usam *word embedding*, que consegue uma representação dos significados das palavras através de vetores dimensionais e, o *Concept Net* que encontra palavras relacionadas com a pesquisa feita.

A. *Preparação e tratamento geral dos dados*

Nesta fase de preparação e tratamento geral dos dados, vamos enumerar os procedimentos tidos em conta para o nosso conjunto de dados bem como a alteração tida em conta. Este tratamento é comum a todas as fases do projeto.

1. Pontuação
   A pontuação está presente em praticamente todos os *reviews* e, de forma a melhorar o tratamento, decidimos remover a pontuação e quebras de página.

2. Conversão Maiúsculas para Minúsculas
   Para evitar que a mesma palavra, escrita em maiúsculas ou minúsculas, seja caracterizada de maneira diferente, por exemplo "*Car*" e "*car*", decidimos converter todos os caracteres para minúsculas.

3. *Stop Words*
   Para melhor análise do texto das *reviews* removemos todas as *stop words* encontradas.

4. Números
   Todas as referências a números foram removidas.

5. Números cardinais e ordinais
   Em complemento com a remoção acima mencionada, decidimos remover também todos os números cardinais e ordinais, por exemplo, *"one"*, *"first"*, etc.

6. Palavras com apenas um caracter
   Todas as palavras que contivessem apenas um caracter foram removidas.

B. *Tratamentos complementares*

Os tratamentos complementares são realizados em conjunto com o tratamento geral. Estes tratamentos são utilizados tanto na fase de Representação de documentos como na Inferência. São abordagens distintas entre si, com o objetivo de entender se a sua implementação tem alguma influência nos resultados, e se tiver, qual ou quais obtém melhores resultados. Podem ser utilizados conjuntos destes tratamentos entre si.

1. *Lemmatization*
   É realizada a lematização nas palavras de cada *review*.

2. *Part-of-speech (POS) tagging*
   Com esta abordagem, o objetivo foi "filtrar" para cada *review* e obter apenas os nomes, resultando assim em *reviews* apenas com nomes.

3. Noun Phrase Chunking
   Abordagem vulgarmente utilizada neste tipo de problemas, realizando o agrupamento de palavras tendo em conta o seu significado, visto que existem palavras que apenas agrupadas apresentam o seu verdadeiro significado. Para esta abordagem utilizou-se a biblioteca *spacy*.

Após o tratamento geral e antes de ser criado os nossos modelos de tópicos, foi criado um dicionário com o conjunto de dados de treino e também foi criado uma *document-term matrix* através desse dicionário. Este dicionário e matriz foram criados com o objetivo de serem utilizados no treino dos modelos propostos.

IV. CONSTRUÇÃO DOS MODELOS DE TÓPICOS

Nesta etapa vamos construir os nossos modelos de tópicos, mais concretamente o modelo *LDA* e o modelo *LSA*. Para ambos os modelos são definidos o número de tópicos e o número de palavras associado a cada tópico. Para o número de tópicos foi utilizado 8, por ser o número de categorias identificadas anteriormente e como número de palavras foi utilizado 15, por ser um número suficientemente grande para que a utilização das bibliotecas seja feita com maior precisão. Recorremos à biblioteca *gensim* para a criação dos modelos referidos, na qual os parâmetros passados foram a *document-term matrix*, o dicionário e o número de tópicos. Apenas para o modelo *LDA* foi passado também o parâmetro de número de passagens, que no nosso caso foram utilizadas 40 passagens, para ser possível normalizar os resultados.

Nas 2 tabelas seguintes, uma para o modelo *LSA* e outra para o modelo *LDA*, são mostrados para cada tópico o nome do tópico caracterizado através das bibliotecas e as palavras que representam cada um dos tópicos.

[1] https://relatedwords.org/

1. Tabela Correspondência de Tópicos – *LSA*

Tópico	Caracterização	Palavras
0	*Computers*	*'like','car', 'also', 'would', 'get', 'even', 'imac', 'good', 'much', 'well', 'time', 'dell', 'new', 'really', 'system'*
1	*Computers*	*'car', 'imac', 'dell', 'mac', 'pc', 'apple', 'computer', 'rear', 'photos', 'applications', 'seat', 'mail', 'iphoto', 'mouse', 'engine'*
2	*Music*	*'car', 'track', 'song', 'beat', 'album', 'lyrics', 'dell', 'like', 'system', 'drive', 'murphy', 'chorus', 'hip', 'spits', 'hop'*
3	*Computers*	*'dell', 'imac', 'system', 'mac', 'costumer', 'apple', 'pc', 'software', 'care', 'photos', 'may', 'iphoto', 'applications', 'order', 'dtc'*
4	*Cookware*	*'clad', 'pan', 'stainless', 'steel', 'pans', 'song', 'car', 'track', 'cookware', 'fry', 'album', 'set', 'beat', 'room', 'lyrics'*
5	*Cookware*	*'clad', 'stainless', 'pan', 'room', 'steel', 'hotel', 'pans', 'fry', 'disney', 'resort', 'car', 'movie', 'cookware', 'song', 'film'*
6	*Phones*	*'phone', 'room', 'handset', 'panasonic', 'hotel', 'battery', 'base', 'resort', 'disney', 'phones', 'cordless', 'handsets', 'caller', 'beach', 'pool'*
7	*Movies*	*'phone', 'movie', 'film', 'room', 'hotel', 'book', 'story', 'plot', 'resort', 'santa', 'also', 'handset', 'disney', 'panasonic', 'beach'*

2. Tabela Correspondência de Tópicos – *LDA*

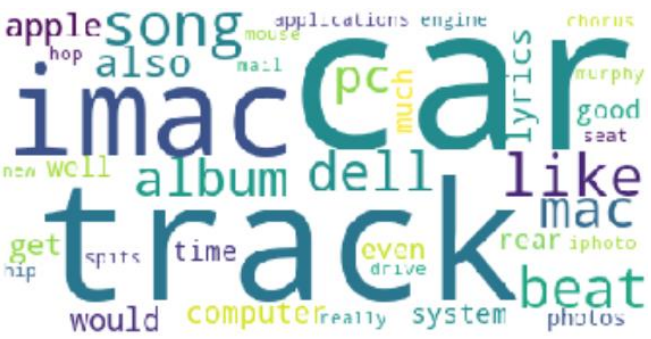
Tópico	Caracterização	Palavras
0	*Computers*	*'computer', 'apple', 'dell', 'imac', 'system', 'also', 'like', 'pc', 'drive', 'new', 'mac', 'machine', 'get', 'would', 'dvd'*
1	*Cookware*	*'pan', 'pans', 'stainless', 'cookware', 'clad', 'set', 'steel', 'use', 'heat', 'like', 'non', 'pots', 'stick', 'dishwasher', 'pot'*
2	*Cars*	*'car', 'like', 'cl', 'type', 'bmw', 'also', 'engine', 'acura', 'drive', 'much', 'driver', 'would', 'rear', 'well', 'seats'*
3	*Cars*	*'car', 'like', 'engine', 'ford', 'rear', 'also', 'seat', 'much', 'cars', 'good', 'front', 'power', 'well', 'would', 'driving'*
4	*Phones*	*'phone', 'handset', 'phones', 'panasonic', 'system', 'like', 'good', 'base', 'use', 'handsets', 'cordless', 'great', 'ghz', 'also', 'would'*
5	*Movies*	*'movie', 'film', 'like', 'plot', 'even', 'cat', 'santa', 'make', 'would', 'get', 'see', 'details', 'good', 'story', 'time'*
6	*Hotels*	*'room', 'hotel', 'resort', 'disney', 'time', 'stay', 'rooms', 'pool', 'also', 'like', 'would', 'beach', 'get', 'night', 'us'*
7	*Music*	*'book', 'album', 'track', 'like', 'song', 'beat', 'story', 'lyrics', 'would', 'time', 'get', 'read', 'good', 'even', 'hip'*

3. *Word Cloud*

Nesta etapa é criada uma *Word Cloud* para cada um dos modelos construídos anteriormente (LDA e LSA), tendo em conta os três tópicos mais importantes para cada um dos modelos. Estas *Word Clouds* foram criadas através da biblioteca *wordcloud*. Elas são criadas através das palavras obtidas em cada modelo, recorrendo aos seus pesos. Esta técnica é importante para ser possível visualizar as palavras com maior peso ou importância para cada modelo. A importância das palavras reflete-se no tamanho de cada palavra, em que quanto maior a palavra, maior importância ela tem para determinado modelo.

A. WordCloud LSA

No modelo LSA podemos verificar que as palavras com mais destaque são *"car", "imac"* e *"track":*



B. WordCloud LDA

Contudo no modelo LDA podemos verificar que as palavras com mais destaque são *"movie", "car", "book", "álbum"* e *"film"*.

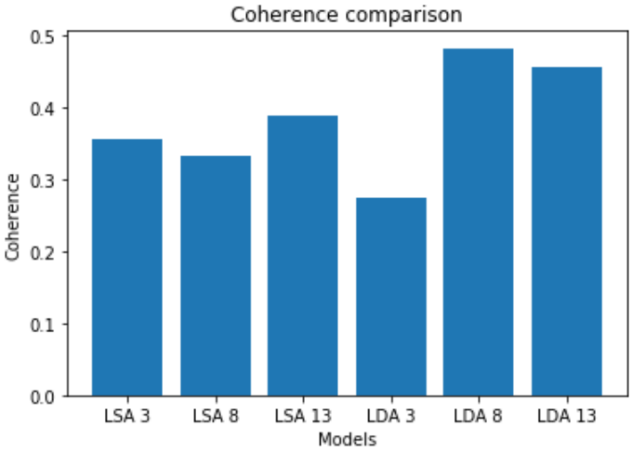
4. Análise dos modelos anteriores

A caracterização dos tópicos dos modelos anteriores já foi realizada de forma automática pela aplicação do método que utiliza as bibliotecas mencionadas. Essa caracterização está espelhada nas 2 últimas tabelas. Num modo geral, analisando a categorização realizada pelo método que utiliza as bibliotecas, confirma-se que esta foi realizada corretamente, apesar de existir alguma ambiguidade no que toca às palavras de determinado tópico, sendo que poderia ter categorizado de forma diferente. Esta distinção entre uma categoria e outra é feita com base no peso das palavras de cada tópico, sendo que para uma determinada categoria é somado os pesos das palavras presentes nesse tópico e que se encontram na biblioteca associada e no final é escolhida a categoria com maior número de somatório de pesos por tópico.

5. Número de tópicos variados

Desta fase em diante, iremos utilizar a métrica *coherence*, para ser possível avaliar de forma mais objetiva os resultados, i.e., a comparação de *performance* entre os modelos utilizados. Esta métrica, muito resumidamente, mede a coerência de um tópico, calculando a concorrência de cada palavra do tópico com todas as palavras desse tópico.

De seguida, repetimos o processo alterando o número de tópicos para 3 e 13. No gráfico abaixo conseguimos comparar os diferentes valores de *coherence* consoante os diferentes números de tópicos que era passado ao modelo.



De acordo com o gráfico, verificamos que para o modelo *LSA*, usando 13 tópicos e para o modelo *LDA*, a utilização de 8 tópicos gerou melhores resultados de *coherence*, sendo que o *LDA* com 8 tópicos foi o modelo que obteve *coherence* mais elevada. No caso de números demasiado baixos ou demasiado altos de tópicos, tornou-se difícil de categorizar, no entanto, o método com as bibliotecas conseguiu apresentar categorizações corretas, comparando com uma caracterização manual.

V. REPRESENTAÇÃO DOS DOCUMENTOS

Após os resultados obtidos, foram realizadas mais e diferentes experiências, com o objetivo de melhorar a representação dos documentos. Como já mencionado os modelos utilizados nas fases anteriores foram o *LSA* e o *LDA*, e nesta fase para além da aplicação destes modelos, irá ser feita a aplicação do modelo *HDP* (*Hierarchical Dirichlet Process*). Este modelo, como os demais, foi construído a partir da biblioteca *gensim*, e este particularmente foi criado apenas com a *document-term matrix* e o dicionário.

Como já mencionado, foram utilizados tratamentos complementares, de forma a melhorar a representação dos documentos. No caso da utilização do *Stemming*, decidimos rejeitá-la porque logo à partida era sabido que seria uma abordagem que não iria trazer melhorias e no máximo iria manter os mesmos resultados, isto porque iria originar palavras que não existem e no caso do método com as bibliotecas não seria possível encontrar essas palavras.

Mais uma vez iremos utilizar a métrica coherence como medida de comparação. Os resultados estão presentes na seguinte tabela.

Tratamento	LSA	LDA	HDP
TG	0.33	0.51	0.34
TG +*Lemmatization*	0.32	0.45	0.36
TG + POS	0.33	**0.66**	0.43

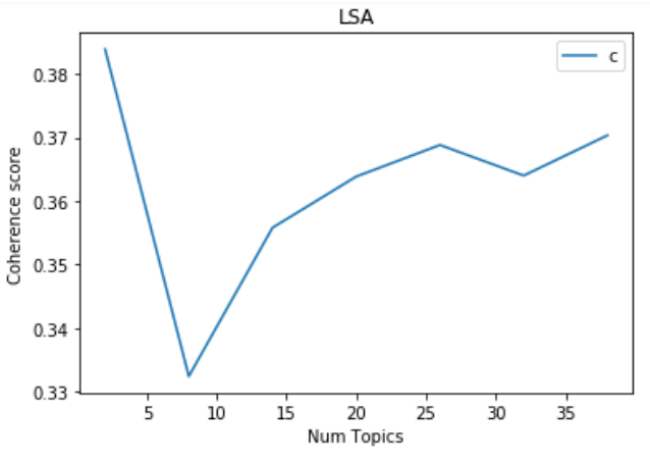
TG +*Chunking*	**0.41**	0.46	0.52
TG +*Lemmatization* + *Chunking*	0.33	0.60	0.46
TG + POS +*Chunking*	0.36	**0.66**	**0.51**
TG +*Lemmatization*+ POS +*Chunking*	0.34	0.56	0.46

Os resultados obtidos foram claramente superiores ao previsto, principalmente com a utilização das abordagens de Tratamento geral com *POS*, Tratamento geral com *Chunking* e ambas em conjunto. Para o *LSA* a melhor abordagem foi Tratamento geral com *Chunking*, sendo que para o *HDP*, a melhor abordagem foi a utilização de Tratamento geral com *POS* e *Chunking*. O *LDA* foi o modelo que apresentou melhores resultados, tendo tido valores bem superiores ao esperado, como já mencionado, na qual as melhores abordagens foram o Tratamento geral com *POS* e o Tratamento geral com *POS* e *Chunking*. Claramente que o fator diferenciador foi a abordagem *POS*.

VI. NÚMERO IDEAL DE TÓPICOS

Este capítulo foca-se numa análise extra que decidimos explorar, onde através da *coherence* e do número de tópicos quisemos descobrir com quantos tópicos, num intervalo de 0 a 35 tópicos, conseguíamos obter o valor máximo de *coherence* para cada um dos modelos usados ao longo do trabalho. Utilizámos para isto o tratamento com valores mais elevados de *coherence*. Nas imagens em baixo é possível observar a evolução do *coherence* à medida que o número de tópicos aumenta.

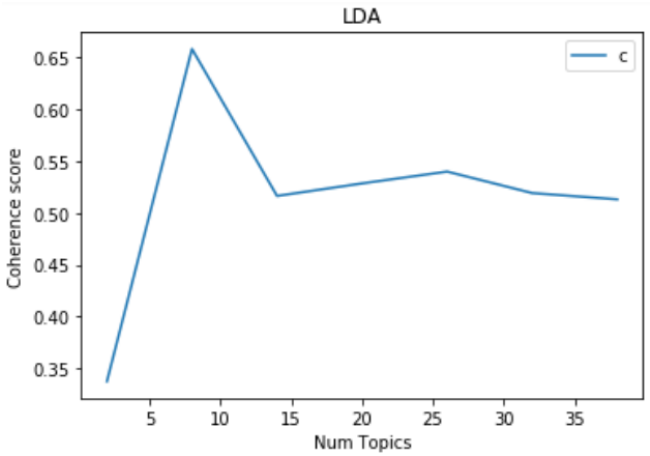
LSA



Para o modelo *LSA* podemos observar que o valor máximo de *coherence* é obtido logo no início com um número de tópicos inferior a 5 e um valor superior a 0.38 de *coherence.*

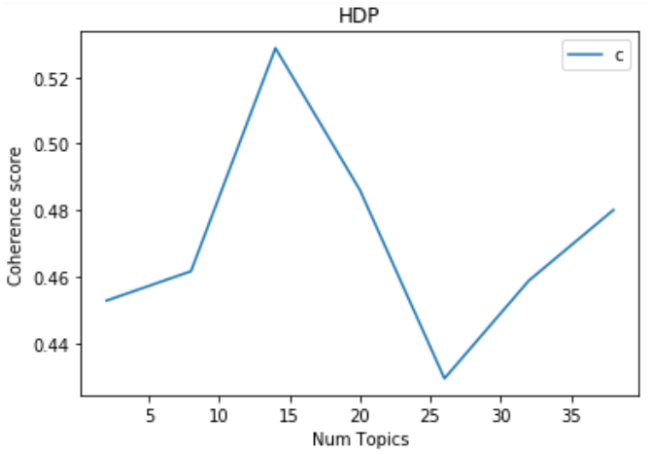
Outra conclusão que podemos tirar é que o valor mais baixo de *coherence* regista-se com aproximadamente 8 tópicos, que é o número de categorias identificadas inicialmente e a partir do momento que temos mais de 10 tópicos a *coherence* aumenta gradualmente.

LDA



No modelo *LDA* podemos observar que até ao número de tópicos 15 a linha do gráfico é oposta à do modelo LSA. O valor máximo de *coherence* é obtido quando temos aproximadamente 8 tópicos. Apesar de descer abruptamente o valor de *coherence* até aos 15 tópicos, depois consegue estabilizar não diferenciando muito o valor de *coherence* à medida que o número de tópicos aumenta.

HDP



Por fim o modelo *HDP* é o que tem maiores oscilações do valor de *coherence* obtendo o seu valor máximo de mais de 0.52 aos 15 tópicos e o seu valor mínimo a rondar os 26 tópicos e menos de 0.44 de *coherence*.

VII. INFERÊNCIA

Neste capítulo, vamos explicar o processo utilizado para a inferência, onde usamos os últimos 10 documentos do conjunto de dados como foi explicado no capítulo da preparação dos dados. Nestes 10 documentos vamos fazer uma categorização manual, ou seja, analisar o documento e pelo seu conteúdo, entender em qual das categorias

mencionadas se enquadra. Na tabela abaixo mostra-se quais foram as conclusões obtidas desta análise.

Documento	Categoriazação
0	Books
1	Phones
2	Movies
3	Music
4	Books
5	Music
6	Computers
7	Music
8	Books
9	Music

Após a leitura e categorização, vamos aplicar os 7 tratamentos explicados no capítulo 5 e analisar qual o comportamento dos três modelos tendo em conta o modo como a categorização foi feita e com isto, conseguimos ter um método de comparação entre experiências.

Com base nos pesos dos tópicos do conjunto de dados de Inferência, iremos averiguar qual o id do tópico que contém maior valor e de seguida com base no resultado da categorização do capítulo 5, é realizada uma busca por esse id que nos devolve o tópico caracterizado.

Tratamento	*LSA*	*LDA*	*HDP*
Tratamento Geral (TG)	5/10	1/10	1/10
TG + *Lemmatization*	5/10	4/10	0/10
TG + ***POS***	**7/10**	**5/10**	0/10
TG + *Chunking*	**7/10**	0/10	1/10
TG + *Chunking* + *Lemmatization*	6/10	0/10	0/10
TG + *POS* + *Chunking*	2/10	0/10	1/10
TG + *POS* + *Chunking* + *Lemmatization*	2/10	1/10	1/10

Através da análise da tabela acima, concluímos que os melhores resultados foram obtidos em três situações, a junção do modelo *LSA* aplicando a abordagem de Tratamento geral com *POS*, junção do modelo *LSA* com a abordagem de Tratamento geral com *Chunking* ambos com um acerto de 7 categorizações em 10 e, por fim, junção do modelo *LDA* com a abordagem de Tratamento geral com *POS* onde 5 de 10 categorizações se encontravam corretas. Mais uma vez de realçar que a abordagem diferenciadora para a obtenção de melhores resultados foi o *POS*.

De notar que o modelo *HDP* teve valores bastante baixos ou nulos em todas as abordagens de tratamento que apresentámos e com isto, concluímos que é um modelo muito menos apropriado em comparação com os restantes.

VIII. CONCLUSÃO

Tendo em conta a realização de várias abordagens para o processamento dos dados e respetivo tratamento, aplicando 3 modelos distintos, tais como *LSA*, *LDA* e *HDP*, ficou evidente que diferentes abordagens dão origem a resultados dispares. Concluímos então, que a escolha da melhor abordagem para o tratamento dos dados é de extrema importância, se o objetivo for a obtenção de melhores resultados. Neste sentido, a abordagem que permitiu obter melhores resultados foi o *Part-of-speech tagging* (POS), tanto em termos da métrica *coherence* como também na Inferência.

Em relação aos modelos, o *LDA* foi o modelo que apresentou melhores resultados em termos da métrica *coherence*, i.e., foi o modelo que mais consistentemente foi capaz de identificar tópicos coerentes, não repetindo muitos tópicos, mas em termos de Inferência não apresentou bons resultados, especialmente quando aplicada a abordagem de *Chunking*. O *LSA* ao contrário do *LDA*, foi o modelo com pior métrica *coherence*, com tendência a demonstrar mais tópicos iguais, mas foi o melhor modelo em termos de Inferência. Isto pode se dever ao facto de, como já mencionado, o modelo ter tendência a demonstrar mais tópicos iguais e na Inferência foram identificados manualmente as últimas 10 *reviews* com alguns tópicos iguais entre eles. O *HDP* obteve resultados razoáveis em relação à *coherence*, estando entre o *LSA* e o *LDA*, mas na Inferência demonstrou ser completamente ineficaz.

As bibliotecas importadas provaram ser uma ótima técnica para identificar tópicos automaticamente sem intervenção humana, contudo podem apresentar algumas falhas se as palavras não estiverem presentes nessas mesmas bibliotecas.

Como trabalho futuro seria de grande valia melhorar as bibliotecas, utilizar o *LDA* com *z-labels* do *Topic-in-Set Knowledge* [4] e aplicar estes modelos a diferentes e maiores conjuntos de dados, de forma a comparar com os resultados obtidos nestas experiências.

IX. CONTRIBUIÇÃO PARA O TRABALHO

- Patricia Santos, 64732: 40%
- Tiago Almeida, 69362: 60%

X. REFERÊNCIAS

[1] C.-L. Liu, W.-H. Hsaio, C.-H. Lee, G.-C. Lu, e E. Jou, «Movie Rating and Review Summarization in Mobile Environment», *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 42, n. 3, pp. 397–407, Mai. 2012.

[2] D. M. Blei, A. Y. Ng, e M. I. Jordan, «Latent Dirichlet Allocation», p. 8.

[3] V. Carrera-Trejo, G. Sidorov, S. Miranda-Jiménez, M. M. Ibarra, e R. C. Martínez, «Latent Dirichlet Allocation complement in the vector space model for Multi-Label Text Classification», vol. 6, n. 1, p. 13, 2015.

[4] D. Andrzejewski e X. Zhu, «Latent Dirichlet Allocation with topic-in-set knowledge», em *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing - SemiSupLearn '09*, Boulder, Colorado, 2009, pp. 43–48.

[5] I. Titov e R. McDonald, «Modeling online reviews with multi-grain topic models», em *Proceeding of the 17th international conference on World Wide Web - WWW '08*, Beijing, China, 2008, p. 111.