Modelagem de Tópicos com LDA

Pedro Vítor Quinta de Castro

- 14 anos de experiência em desenvolvimento
- Especialização em Machine Learning, 2015-2017
- Mestrado em andamento em Processamento de Linguagem Natural
- Desenvolvimento de trabalhos com LDA em motores de busca

Roteiro

- O que é LDA?
- LDA como aprendizado não supervisionado
- Pré-processamento dos textos
- Visualização de Tópicos
- Coerência de Tópicos

O que é LDA?

Definição por exemplificação

- Gosto de comer cenouras e ovos
- Comi ovos com suco de laranja de café da manhã
- Chinchilas e gatinhos são fofos
- Minha irmã adotou um gatinho ontem
- Olhe para este hamster fofo roendo pedaços de cenouras

LDA (Latent Dirichlet Allocation) é uma forma automática de descobrir **tópicos** que estas frases contém.

Descobrindo 2 tópicos

- Frases 1 e 2: 100% tópico A
- Frases 3 e 4: 100% tópico B
- Frase 5: 60% tópico A e 40% tópico B
- Tópico A: 30% ovos, 15% cenoura, 10% café da manhã, 10% roendo, ... (podemos interpretar o tópico A como sendo sobre comida)
- Tópico B: 20% chinchilas, 20% gatinhos, 20% fofo, 15% hamster, ... (podemos interpretar o tópico B como sendo sobre animais fofos)

Como funciona?

LDA representa documentos como sendo uma mistura de tópicos que gera palavras com certas probabilidades

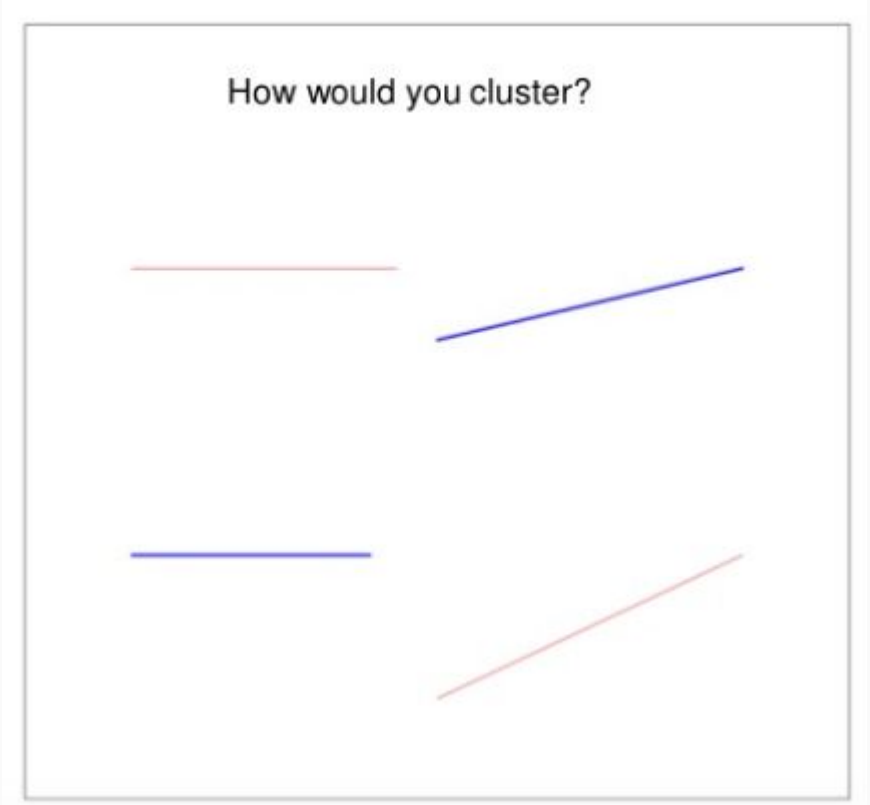
Assume que documentos são produzidos da forma ao lado...

- Decidir o número N de palavras que o documento terá
- Escolher a mistura de tópicos que o documento terá
  - Por exemplo, 33% comida e 67% animais fofos
- Cada palavra do documento é gerada:
  - Escolhe-se o tópico de acordo, de tal forma que a probabilidade de um tópico ser escolhido é correspondente à proporção dele no documento
  - Usa o tópico para escolher as palavras, de acordo com a probabilidade de cada uma (30% de chance de escolher "ovos" se for o tópico A, 20% de chance de escolher "chinchilas" se for o tópico B, etc...)

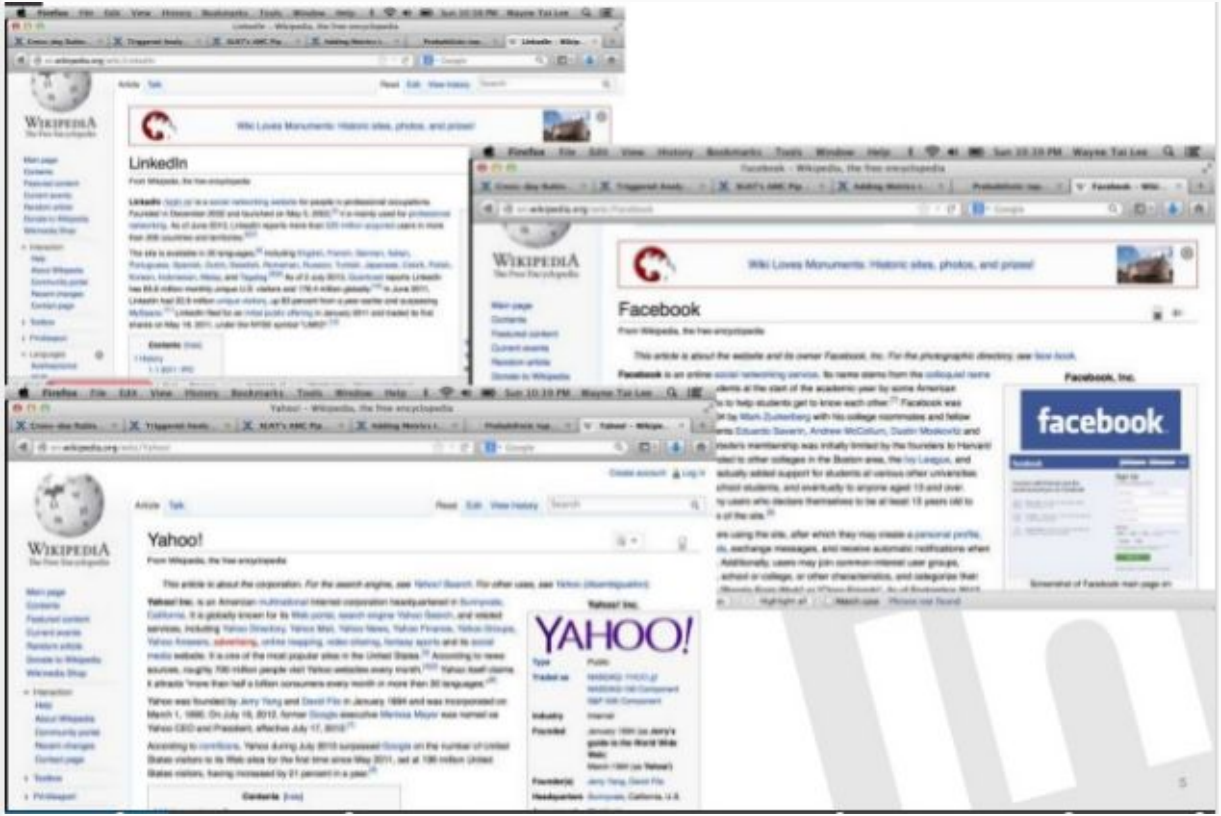
Ir para notebook lda_training_tips

LDA como aprendizado não supervisionado

Como você agruparia os dados ao lado?



E estes?



LDA como Clusterizador

- LDA implementa a clusterização não supervisionada de documentos não estruturados
- Clusterização não é feita a partir de uma medida de similaridade (como distância euclidiana) entre os documentos, mas a partir de inferências calculadas a partir de um modelo de estatística bayesiana
- Além da clusterização e categorização automática dos documentos, também consegue-se a extração automática de tags, a partir das palavras que compõem os tópicos inferidos

Ir para notebook lda_training_offline

Pré-processamento dos textos

“Garbage In, Garbage Out”

Métodos

- Remoção de pontuação, acentos...
- Lowercase
- Remoção de stopwords
- Remoção de "lixo" (geralmente necessário após aplicação de OCR)
- Remoção de termos raros ou muito frequentes
- Remoção de palavras por tamanho (grandes demais ou pequenas demais)
- Remoção de palavras por classe gramatical
    - Aplicação de Part of Speech (POS) Tagging

Métodos

- Detecção de n-grams
  - Tokens com co-ocorrência frequente. Normalmente é uma primeira tentativa de encontrar uma estrutura oculta no corpu. Normalmente bigrams, podendo até trigrams.
- Stemming
  - Redução das palavras à sua raiz. Geralmente é preferível se os dados não serão exibidos.
- Lemmatization
  - Redução das palavras à uma forma substantivada, primitiva. Ignora tempo verbal, gênero ou plural. Geralmente é preferivel ao stemming, na modelagem de tópicos, já que as palavras permanecem compreensíveis.

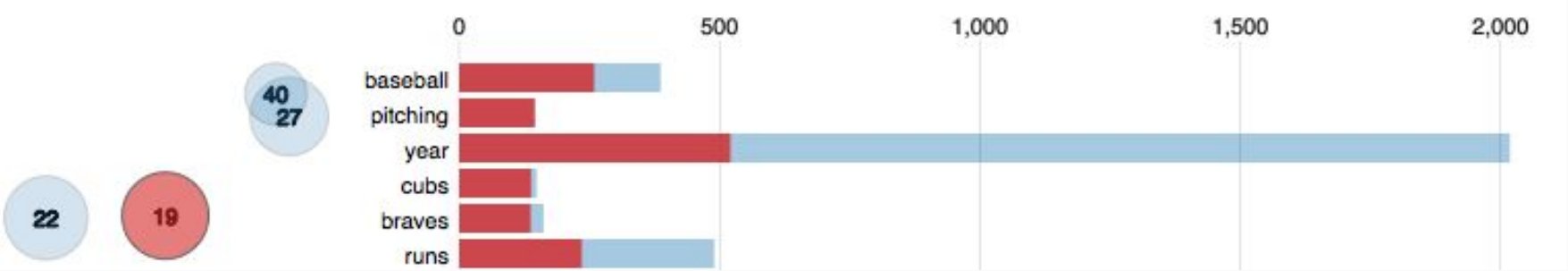
Ir para notebook text_pre_processing

Ir para notebook exercicios_1

Visualização de Tópicos

pyLDAvis

- Ferramenta interativa para visualização de modelagem de tópicos
- Visualização no jupyter notebook
- Exibe os tópicos de acordo com métricas de saliência e relevância

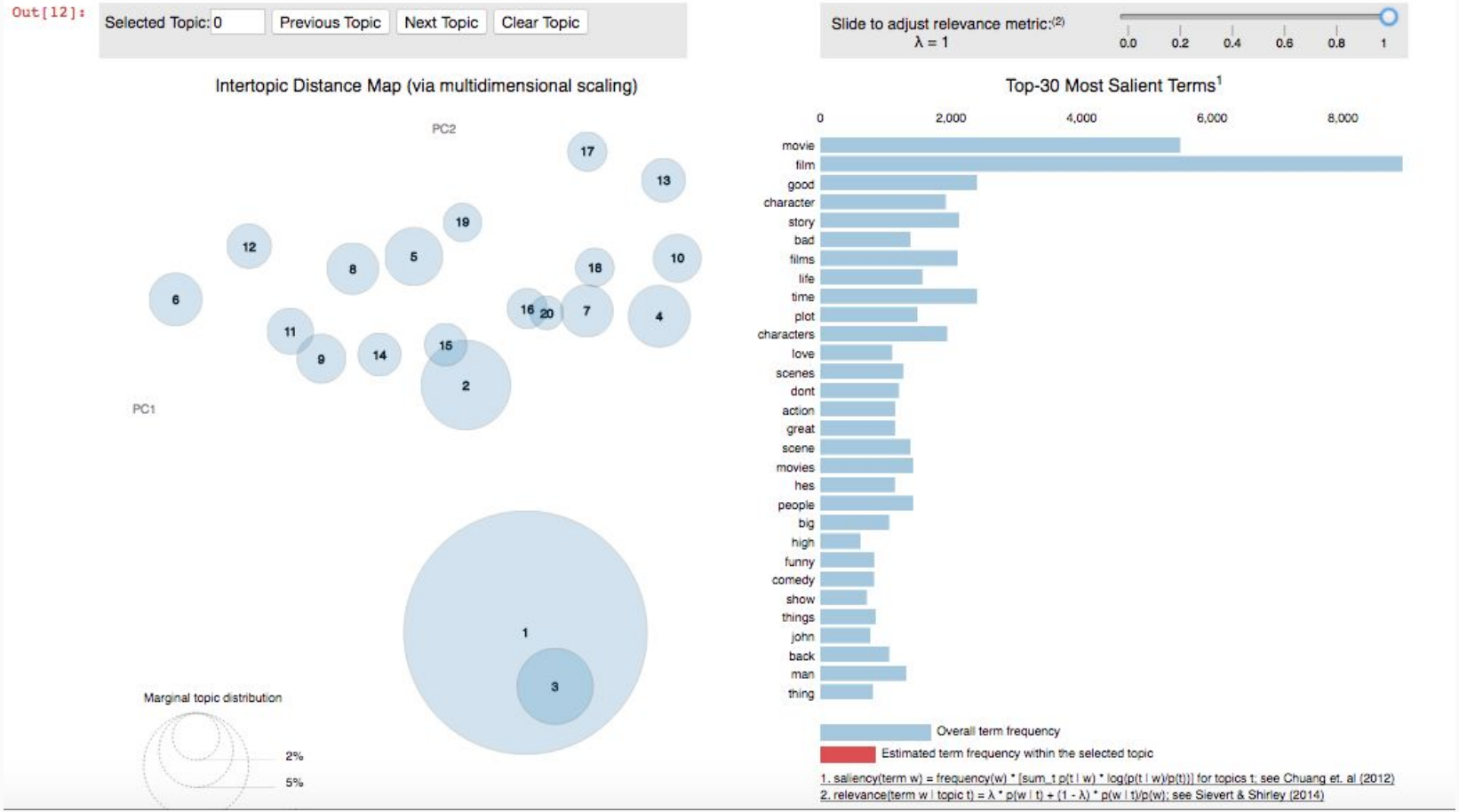


Medidas visualizadas

- Saliência
  - Avalia quão informativa uma palavra é para os tópicos inferidos. Por exemplo, uma palavra que é frequente em **todos** os tópicos não é necessariamente informativa, já que não caracteriza nenhum, exatamente.
- Relevância
  - Mede a contribuição de uma palavra em um determinado tópico inferido. É parametrizada para controlar a contribuição de um termo no tópico selecionado e a contribuição do termo no corpus inteiro.

```
In [12]: pyLDAvis.display(movies_vis_data)
```

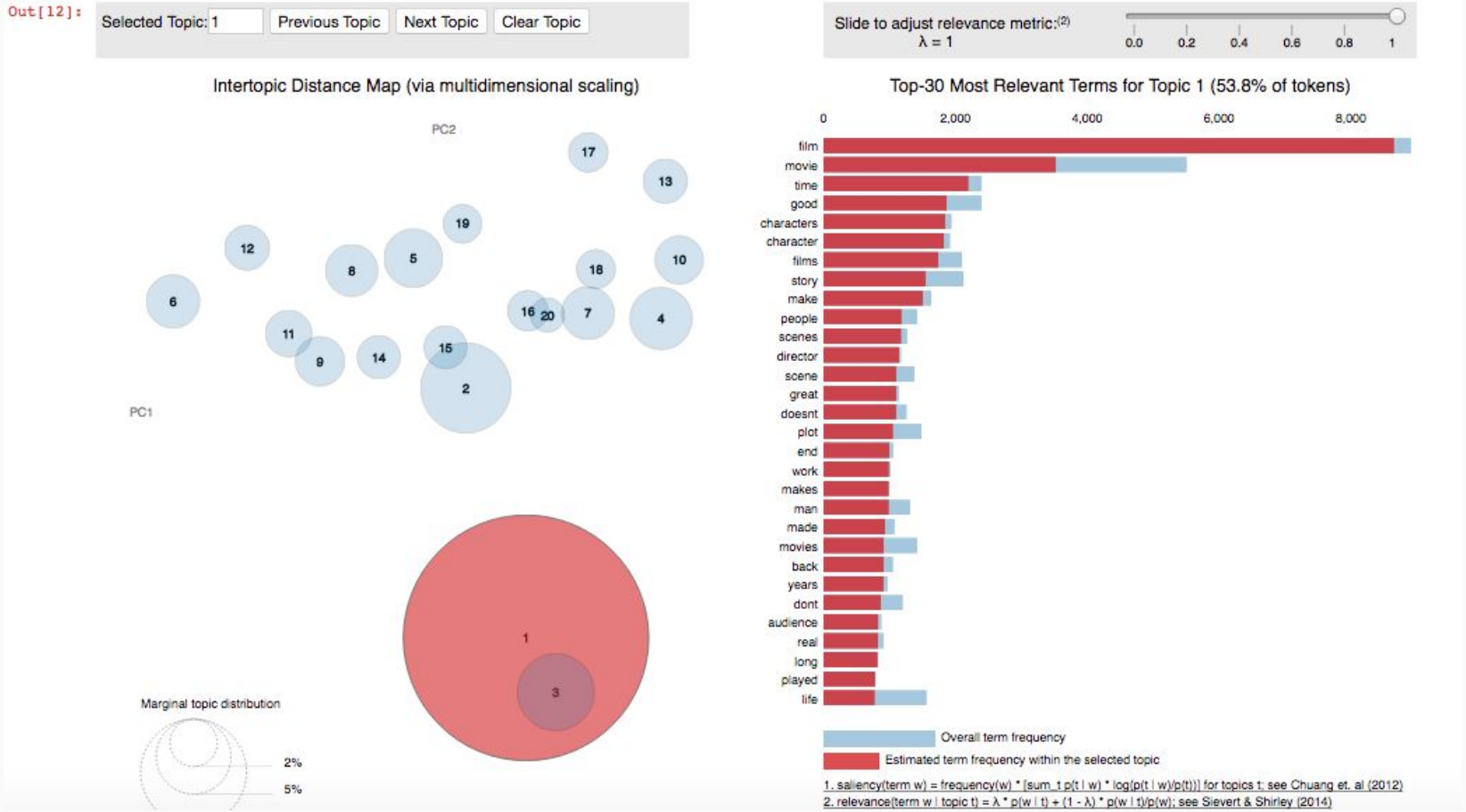
Out[12]:



Visualização com ênfase na saliência dos termos

```
In [12]: pyLDAvis.display(movies_vis_data)
```

Out[12]:



Visualização dos tópicos com ênfase na relevância dos termos para o tópico 1

Ir para notebook topic_visualization

Coerência de Tópicos

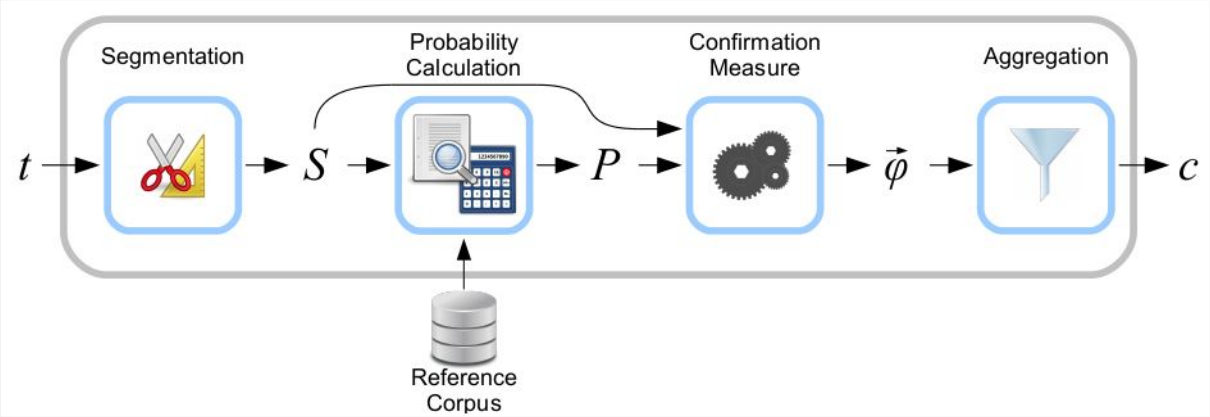
Definição

- Métrica para avaliação da interpretabilidade humana de tópicos
- Critério objetivo e quantitativo para avaliar tópicos obtidos a partir de modelos treinados
- Visa substituir um critério subjetivo e qualitativo de avaliação de tópicos
- Treinar um modelo até que a coerência dos tópicos atingida atinja um valor satisfatório

O Algoritmo

Implementa um pipeline de coerência de tópicos

Este pipeline é um framework em que cada um dos componentes pode ser implementado de forma diferente, provendo diferentes formas de avaliação



- Segmentação
  - As palavras do dicionário são segmentadas de acordo com algum critério
- Cálculo de Probabilidades
  - Define a forma como probabilidades são calculadas a partir dos dados segmentados
- Medida de Confirmação
  - Define uma métrica a partir das probabilidades calculadas e sobre como os segmentos se suportam
- Agregação
  - Agrega as medidas calculadas para produzir um score final

Ir para notebook topic_coherence_tutorial

Ir para notebook exercicios_2

Obrigado!

pvcastro@gmail.com