Suggestion Mining from Online Reviews and Forums

Paulo Henrique da Silva

paulohsilva@inf.ufg.br

UFG
UNIVERSIDADE FEDERAL DE GOIÁS

Novembro, 2018

Processamento de Linguagem Natural

Roteiro

- Introdução
- Trabalhos relacionados
- Processamento de Linguagem Natural
- Problema
- Dados e métodos
- Conclusões

Introdução

- Contexto
  - O aumento no conteúdo on-line influenciou o comportamento de compra dos usuários.
  - O consumidor não é mais influenciado pelos profissionais de marketing
  - É infuenciado pelos comentários dos usuários

Introdução

- Contexto
  - Extração de sugestões de texto não estruturado
  - Opiniões expressam sentimento positivo, negativo ou neutro
  - Sugestões expressam dicas, conselhos ou recomendações
  - Expressas através de revisões on-line: blogs, fóruns de discussões, plataforma de mídias sociais

Introdução

- Exemplo

Introdução

- Motivação
  - Pessoas, empresas e governos querem saber a opinião a respeito de produtos, marcas, serviços ou políticas públicas

Introdução

- Objetivo
  - Desenvolver um modelo automático capaz de extrair informação relevante de revisões on-line através da mineração de opinião
  - Permitir que consumidores extraiam os principais tópicos cobertos pelas revisões
  - Auxiliar empresas a ter *feedback* do consumidor para melhorar seus produtos

Trabalhos relacionados

- Dave et al. [2003] – Opion extraction product review
  - Ferramenta que seleciona e sintetiza análise de produtos
  - *Feature selection and classification*

- Pang et al. [2008] – Opinion mining and sentimental analysis
  - Pesquisa (survey) na área de mineração de opinião e análise de sentimento
  - Identifica as principais tarefas de mineração de opinião e análise de sentimento

Trabalhos relacionados

- Binali et al. [2009] – State of the art opinion mining
  - Avalia os principais trabalhos na aŕea de mineração de opinião

- Sapna Negi [2017] – Suggestion mining from opinionated text
  - Mineração de sugestões e sumarização
  - Utiliza deep learning para classificação

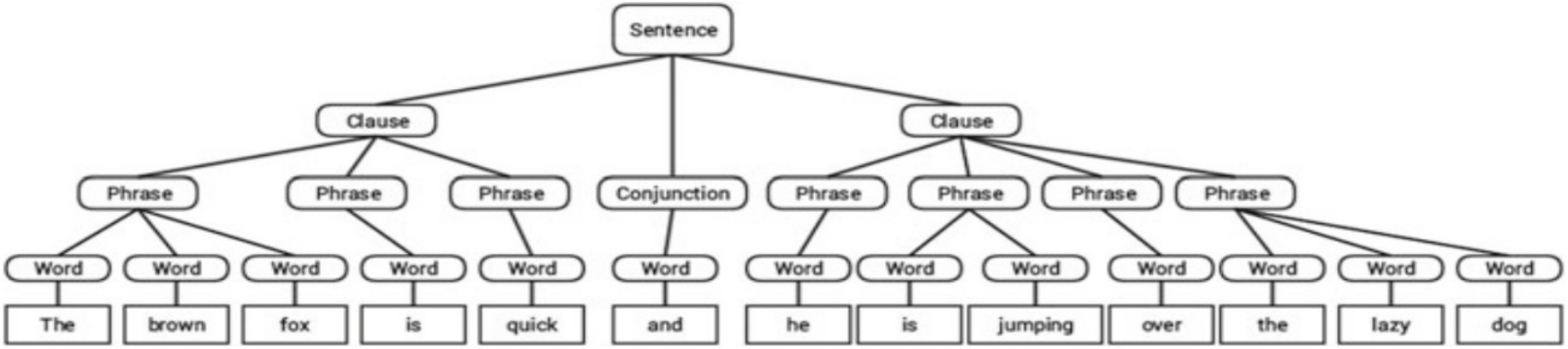
Full suggestion text	Entity	Beneficiary	Keyphrase
If you do end up here, be sure to specify a room at the back of the hotel	Room	Customer	Specify a room at the back of the hotel
If you are here, I recommend a Trabi safari	Trabi Safari	Customer	Trabi Safari
Chair upholstry seriously needs to be cleaned	Chair/Chair upholstry	Brand owner	chair upholstry need to be cleaned

Processamento de Linguagem Natural

- Linguagem natural
  - Desenvolvida pelos humanos ao invés de criado artificialmente
  - Meio de comunicação entre os humanos
  - Espanhol, Inglês, Português
  - Falada, escrita ou sinais

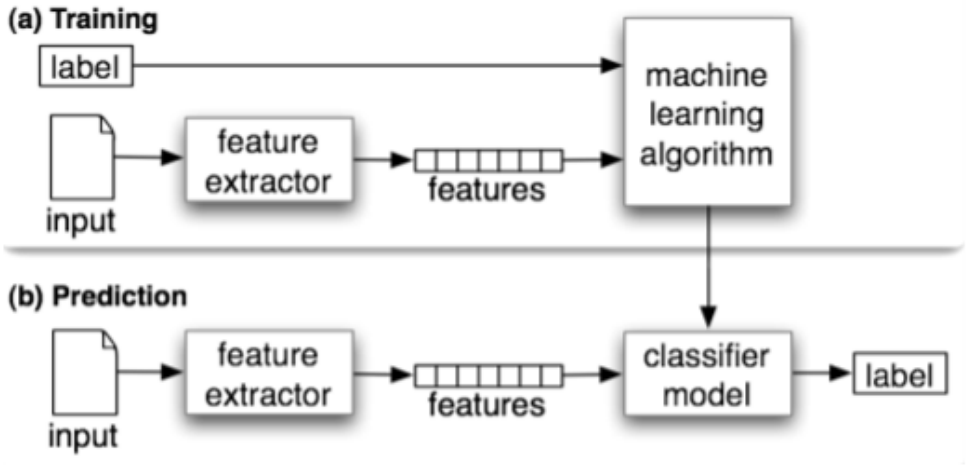
Processamento de Linguagem Natural

- Desafios Linguísticos
  - Dificuldade para a máquina entender a linguagem natural
  - Estrutura e sintaxe, contexto da conversa, gírias, figuras de linguagens etc



Processamento de Linguagem Natural

- Converte linguagem humana em conhecimento que o computador possa entender
- Algoritmos e técnicas aplicadas para extrair conhecimento (insights) dos documentos
- Extração de características sintáticas e semânticas



Processamento de Linguagem Natural

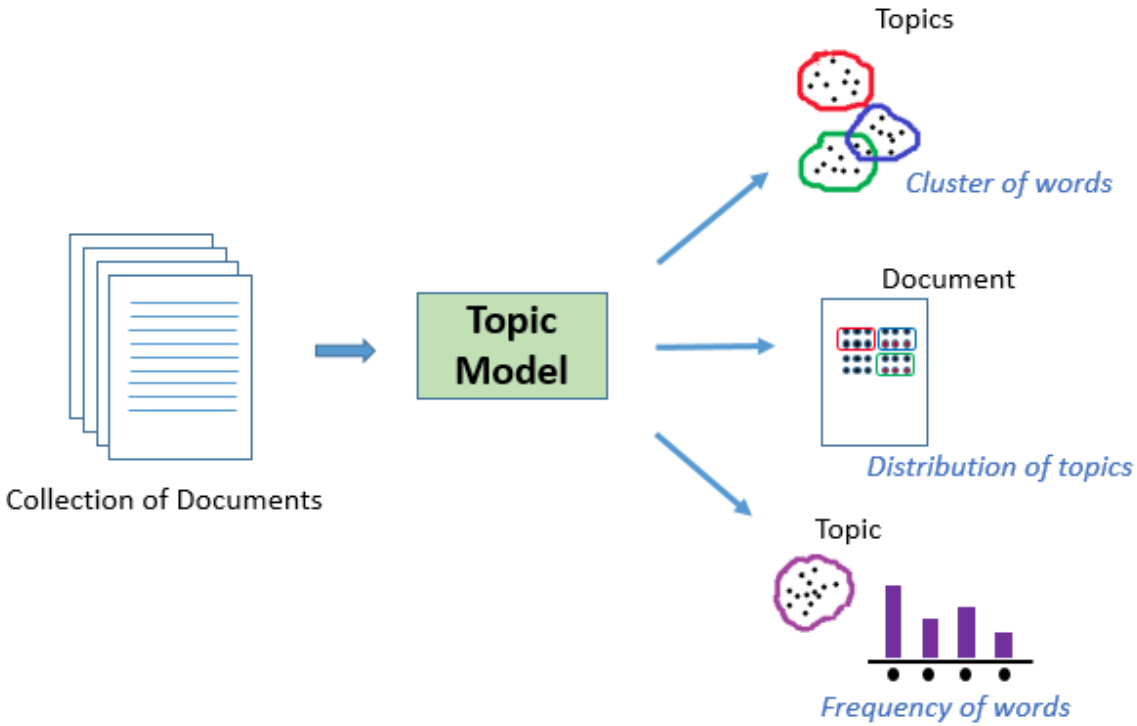
- Aplicações
  - Análise de sentimento – positivo, negativo ou neutro
  - Marcação de classe gramatical – substantivo, verbo, pronome
  - Mineração de sugestões – extração de informação de *feedback*
  - NER – pessoa, local, organização etc
  - Sumarização de textos – resumo de documento ou *corpus*

Problema

- Mineração de sugestões
  - Extrair sugestões de revisões on-line e fóruns
  - Dataset textual
  - Alta dimensionalidade
  - Como analisar um grande n° de revisões on-line usando NLP?

Problema

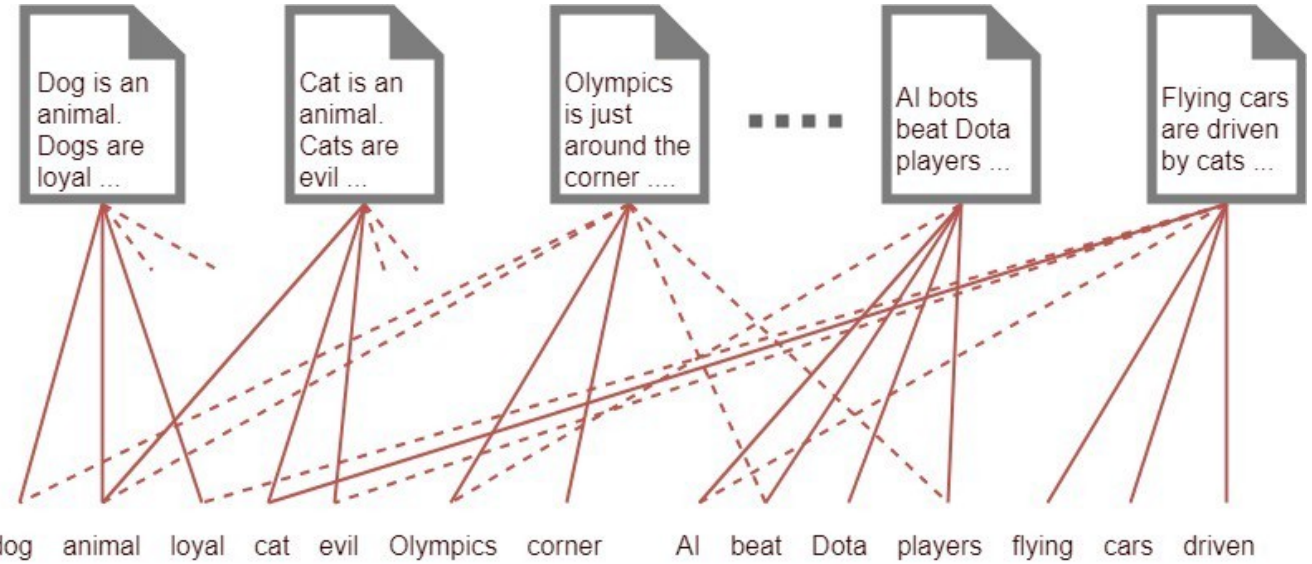
- Modelagem de tópico



https://www.analyticsvidhya.com/blog/2018/10/mining-online-reviews-topic-modeling-lda

Dados e métodos

- Latent Dirichlet Allocation - LDA
  - Documentos conectados pelo conjunto de palavras (tópico)
  - Gera palavras baseadas em sua distribuição de probabilidades



https://towardsdatascience.com/light-on-math-machine-learning-intuitive-guide-to-latent-dirichlet-allocation

Dados e métodos

- Datasets
  - *Suggestion Forums* (Tarefa A) e *Hotel Reviews* (Tarefa B)
  - Anotados em duas fases
    - *Crowdsourced annotators*
    - Especialistas
  - Apenas sentenças que explicitamente expressam sugestões

'I loved the cup cakes from the bakery next door'

is an implicit form of a suggestion which can be explicitly expressed as:

'Do try the cupcakes from the bakery next door'

Dados e métodos

- Pré-processamento dos dados
  - Tokenization
  - Stop-word removal
  - Stemming
  - Lemmatize
  - Bag-of-words
  - TF-IDF

Dados e métodos

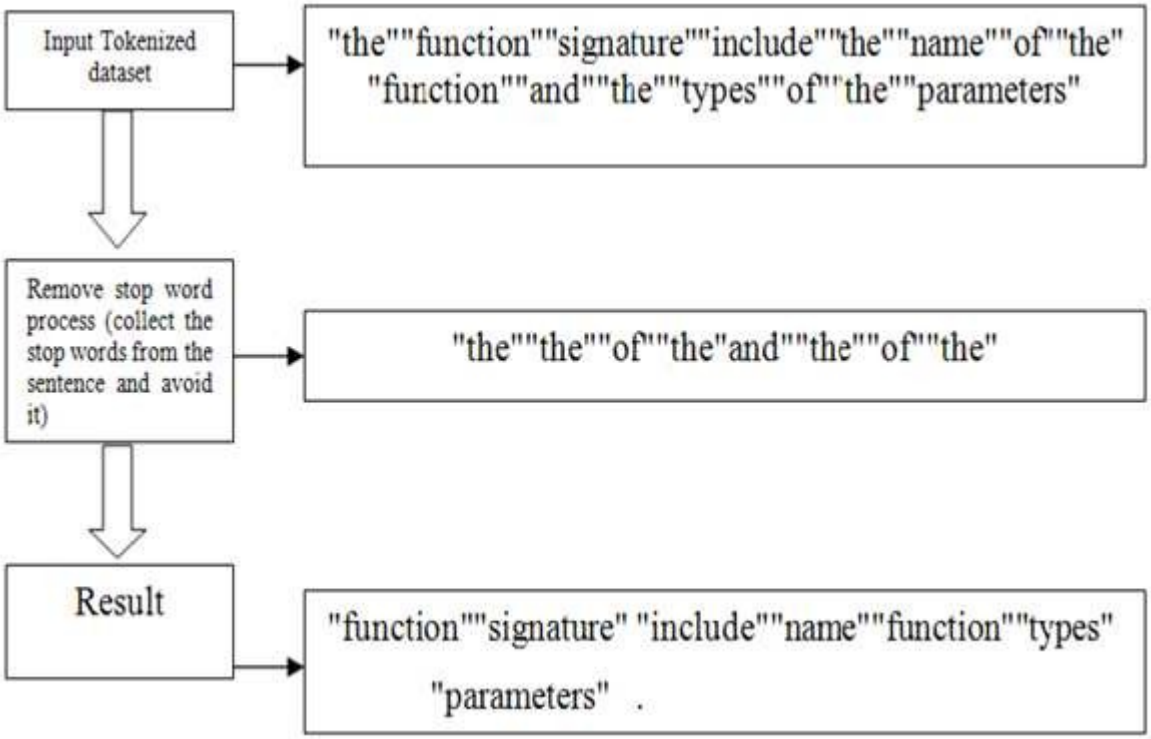
- Tokenization

```
Sentence tokenizer:
['Are you curious about tokenization?', "Let's see how it works!", 'We need to analyze a couple of senten
ces with punctuations to see it in action.']

Word tokenizer:
['Are', 'you', 'curious', 'about', 'tokenization', '?', 'Let', "'s", 'see', 'how', 'it', 'works', '!', 'W
e', 'need', 'to', 'analyze', 'a', 'couple', 'of', 'sentences', 'with', 'punctuations', 'to', 'see', 'it',
 'in', 'action', '.']

Punkt word tokenizer:
['Are', 'you', 'curious', 'about', 'tokenization', '?', 'Let', "'s", 'see', 'how', 'it', 'works', '!', 'W
e', 'need', 'to', 'analyze', 'a', 'couple', 'of', 'sentences', 'with', 'punctuations', 'to', 'see', 'it',
 'in', 'action.']

Word punct tokenizer:
['Are', 'you', 'curious', 'about', 'tokenization', '?', 'Let', "'", 's', 'see', 'how', 'it', 'works', '!'
, 'We', 'need', 'to', 'analyze', 'a', 'couple', 'of', 'sentences', 'with', 'punctuations', 'to', 'see', '
it', 'in', 'action', '.']
```

https://exploreai.org/p/machine-learning-analyzing-text-data

Dados e métodos

- Stop-word removal



https://www.google.com

Dados e métodos

- Stemming

WORD	PORTER	LANCASTER	SNOWBALL
table	tabl	tabl	tabl
probably	probabl	prob	probabl
wolves	wolv	wolv	wolv
playing	play	play	play
is	is	is	is
dog	dog	dog	dog
the	the	the	the
beaches	beach	beach	beach
grounded	ground	ground	ground
dreamt	dreamt	dreamt	dreamt
envision	envis	envid	envis

https://exploreai.org/p/machine-learning-analyzing-text-data

Dados e métodos

- Lemmatizer

WORD	NOUN LEMMATIZER	VERB LEMMATIZER
table	table	table
probably	probably	probably
wolves	wolf	wolves
playing	playing	play
is	is	be
dog	dog	dog
the	the	the
beaches	beach	beach
grounded	grounded	ground
dreamt	dreamt	dream
envision	envision	envision

https://exploreai.org/p/machine-learning-analyzing-text-data

Dados e méodos

- Bag-of-words

Document 1

The quick brown fox jumped over the lazy dog's back.

Document 2

Now is the time for all good men to come to the aid of their party.

Term	Document 1	Document 2
aid	0	1
all	0	1
back	1	0
brown	1	0
come	0	1
dog	1	0
fox	1	0
good	0	1
jump	1	0
lazy	1	0
men	0	1
now	0	1
over	1	0
party	0	1
quick	1	0
their	0	1
time	0	1

Stopword List

for
is
of
the
to

www.google.com

Dados e métodos

- TF-IDF
  - TF – medida de quão importante o termo é para o documento
  - IDF – medida da importância do termo para o *corpus*

$$\text{tfidf} = \text{tf} \cdot \text{idf}$$

$$\text{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

$$\text{idf}_i = \log \frac{|D|}{|\{d : d \ni t_i\}|}$$

www.google.com

- Binary → count → weight matrix

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	5.25	3.18	0	0	0	0.35
Brutus	1.21	6.1	0	1	0	0
Caesar	8.59	2.54	0	1.51	0.25	0
Calpurnia	0	1.54	0	0	0	0
Cleopatra	2.85	0	0	0	0	0
mercy	1.51	0	1.9	0.12	5.25	0.88
worser	1.37	0	0.11	4.15	0.25	1.95

|V|

Each document is now represented by a real-valued vector of tf-idf weights ∈ R|V|

Dados e métodos

- Latent Dirichlet Allocation – LDA
  - Word embedding
  - Feature learning techniques
  - Aplicado nos dados pré-processados

	W1	W2	W3	Wn
D1	0	2	1	3
D2	1	4	0	0
D3	0	2	3	1
Dn	1	1	3	0

Matriz de termo-documento

	K1	K2	K3	K
D1	1	0	0	1
D2	1	1	0	0
D3	1	0	0	1
Dn	1	0	1	0

Matriz de tópico-documento

	W1	W2	W3	Wm
K1	0	1	1	1
K2	1	1	1	0
K3	1	0	0	1
K	1	1	0	0

Matriz de termo-tópico

https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python

Conclusões

- Mineração de opinião baseada em características usando *dataset* de revisões on-line de usuários
- Aplicação de técnicas para identificar e separar sugestões e recomendações dos usuários
- Identificação de relações semânticas e extração de informação relevante
- Uso de *machine learning* para melhorar a acurácia
- Aplicação em várias áreas incluindo entretenimento, política, mercado financeiro e comércio eletrônico

Obrigado!

Suggestion Mining from Online Reviews and Forums

Paulo Henrique da Silva

paulohsilva@inf.ufg.br

UFG
UNIVERSIDADE FEDERAL DE GOIÁS

Novembro, 2018

Processamento de Linguagem Natural