Processamento de Linguagem Natural

Aula 03

Medidas de Similaridade Textual

PUC Minas

Similaridade Textual

Existem várias aplicações para utilizar a similaridade entre textos

- **Engines de busca**, por exemplo sites de pergunta e resposta como o Stackoverflow precisam determinar se uma pergunta já foi respondida antes

Similaridade Textual

- Em **assuntos legais** a tarefa de similaridade textual ajuda a reduzir riscos de contratos e encontrar precedentes de caso, com base em contratos e casos anteriores similares.

Similaridade Textual

- No **atendimento ao consumidor**, IA deve ser apta a entender as dúvidas dos consumidores e dar uma resposta uniforme e coerente.

O que é a Similaridade Textual no PLN?

- Pessoas diferentes tem uma definição diferente para a similaridade de textos
- O objetivo é retornar o quão "próximas" duas partes de um texto são no significado ou na forma da escrita.
  - O primeiro se refere à similaridade semântica
  - O segundo se refere à similaridade léxica

O que é a Similaridade Textual no PLN?

- Algumas vezes os métodos para encontrar similaridade semântica são utilizados para encontrar a similaridade léxica
  - Atingir a total similaridade semântica tem muito mais envolvido.

Similaridade Léxica

- Na maioria das vezes, quando nos referimos à similaridade textual, as pessoas se referem ao quão similares dois textos são na superfície.

- O quão similares as frases "O gato comeu o rato" e "O rato comeu comida de gato" são somente olhando a superfície das palavras?

Similaridade Léxica

- Olhando somente as palavras elas são muito similares, pois ambas compartilham 4 palavras

Intersessão =

O gato comeu o rato

∩

O rato comeu a comida do gato = 4

PUC Minas

Similaridade Léxica

- A noção de similaridade geralmente referência Similaridade Léxica
  - Normalmente não leva em conta o significado das palavras nem as frases para contexto
- Isso não significa que esse tipo de similaridade seja menos efetivo

Similaridade Léxica

- Exemplo: comparando similaridade entre as frases de artigos de jornal.
  - Você pode usar a representação vetorial das palavras ao invés de comparar palavras sozinhas.
  - Ao invés de comparar palavra a palavra você estará dando mais contexto e capturando mais semântica

Granularidade

- Similaridade Léxica pode ser calculada com diferentes granularidades
  - Nível de caracteres
    - Também conhecida como similaridade/proximidade de strings
  - Nível de palavras
  - Nível de frases
    - Quebrando uma parte do texto em grupos de palavras relacionadas antes de calcular a similaridade

Métricas de Similaridade

- Algumas métricas comuns utilizadas para o cálculo de similaridade são:
  - Jaccard
  - Dice
  - Cosine

Similaridade de Jaccard

- Similaridade de Jaccard é definida como o **tamanho da intercessão dividido pelo tamanho da união dos dois conjuntos**
- Utilizaremos dois exemplos em inglês:
  - AI is our friend and it has been friendly
  - AI and humans have always been friendly

$$Jaccard = \frac{|tokens\_in\_string\_A \cap tokens\_in\_string\_B|}{|tokens\_in\_string\_A \cup tokens\_in\_string\_B|}$$

PUC Minas

Similaridade de Jaccard

- Antes de realizar o cálculo iremos realizar o processo de **lemmatização** para reduzir todas as palavras para a mesma raiz:
  - Friend, friendly → friend
  - Has, have → has

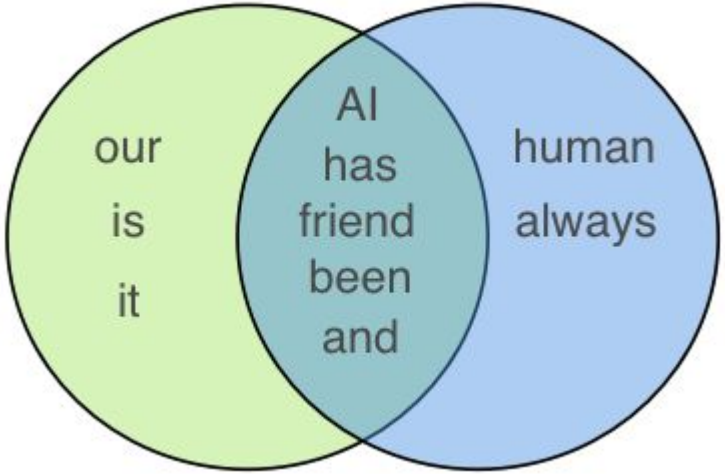
Similaridade de Jaccard

- Similaridade de Jaccard = 0.5

$$J(X,Y) = |X \cap Y| \ / \ |X \cup Y|$$

$$5/(5+3+2) = 0.5$$

A palavra **friend** aparece duas vezes na primeira frase, mas não influenciou no cálculo



Similaridade do Cosseno

- A Similaridade do Cosseno é calculada medindo o **cosseno do ângulo entre dois vetores:**

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

PUC Minas

Similaridade do Cosseno

- Para realizar esse tipo de cálculo precisamos inicialmente transformar as duas sentenças em vetores

- Uma forma é usar **Bag of Words** com **TF**(Term Frequency) ou **TF-IDF**(term frequency- inverse document frequency).
- A escolha entre **TF** ou **TF-IDF** depende da aplicação
  - **TF** é bom para similaridades textuais gerais
  - **TF-IDF** é bom para busca com relevância

Similaridade do Cosseno

- Outra forma é usar **Word2Vec** ou o seu próprio algoritmo de *word embedding.*
  - **Word2Vec** é um par de modelos de aprendizado não supervisionado para criação de uma representação vetorial de palavras presentes em textos que usam linguagem natural.
  - A representação é condicionada à distribuição do texto e apresenta características semânticas.
  - Palavras com significado similar tem vetores próximos e operações aritméticas formam expressões que fazem sentido.

Similaridade do Cosseno

- TF/TF-IDF com Bag of Words
  - Cria um número por palavra
  - Bom para classificação de documentos por inteiro
- Word Embeddings
  - Cria um vetor por palavra
  - Bom para identificar conteúdo contextual

Diferenças entre os métodos de Jaccard e Cosseno

- *Similaridade de Jaccard* utiliza um **conjunto único de palavras** para cada sentença ou documento, enquanto a *Similaridade do Cosseno* utiliza o **tamanho total dos vetores** (que podem ser criados com a frequência de termos do *bag-of-words* ou *tf-idf)*
  - Isso significa que a palavra **friend** repetida mais de uma vez na primeira frase muda o valor da similaridade calculada pelo Cosseno mas não afeta o valor calculado pela similaridade de Jaccard. Caso a palavra **friend** estivesse repetida 50 vezes na primeira frase, a similaridade de Jaccard continuaria sendo 0.5, enquanto a do cosseno cairia para 0.4

Diferenças entre os métodos de Jaccard e Cosseno

- *Similaridade de Jaccard* é boa para casos onde duplicação de palavras não fazem diferença para a análise
- *Similaridade pelo Cosseno* é boa quando as duplicações importam enquanto a análise textual de similaridade está sendo calculada.
    - Para a descrição de dois produtos, a repetição de palavras não reduz a similaridade, então a utilização do cálculo por *Jaccard* seria indicado nesse caso, por ser um método mais simples e rápido.

Quando utilizar Similaridade Léxica?

- Clusterização
  - Quando você precisa agrupar grupos de textos similares juntos, pode ser utilizada para encontrar as similaridades de ambos
- Remoção de Redundância
  - Se duas partes do textos são similares, podemos não precisar de ambos. Pode-se eliminar o que estiver redundante
  - Por exemplo: Lista de produtos, pessoas repetidas em um banco de dados, páginas html duplicadas

Quando utilizar Similaridade Léxica?

- Recuperação de Informação
  - Como rankear documentos que são similares? Podemos usar algo simples como a similaridade de *cosseno*. A medida do Cosseno também pode ser utilizado quando se tem um vetor de representação dos documentos.
  - Pode ser usada a similaridade de Jaccard para tarefas de recuperação de informação, mas não é tão efetiva pela ausência das frequências dos termos.

Similaridade Semântica

- Outra noção de similaridade utilizada pelo PLN é o quão similar é o significado de duas frases?
- Se você olhar as duas frases

O gato comeu o rato

O rato comeu a comida do gato

Grande sobreposição de palavras x Significado completamente diferente

PUC Minas

Similaridade Semântica

- Pegar o significado de frases é uma tarefa mais difícil e que requer um nível mais profundo de análise
- Um aspecto que pode ser observado é a ordem das palavras

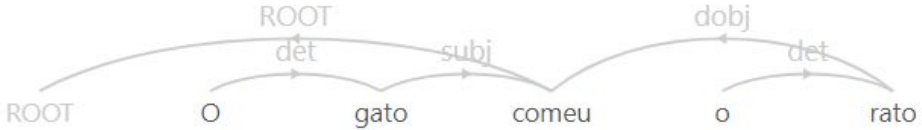
Gato → comeu → rato

Rato → comeu → comida do gato

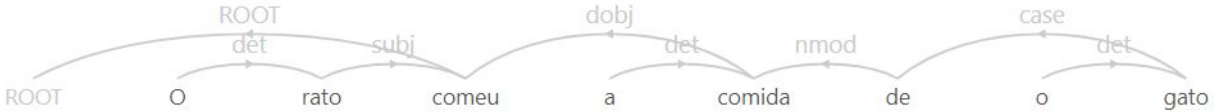
Apesar das frases terem muitas palavras equivalentes, a ordem diferente indica que o significado de ambas é diferente

Similaridade Semântica

- Utilização de análise sintática pode ajudar na similaridade semântica
- Vamos olhar as árvores sintáticas das duas frases



(T) Head position	(T) Head	(T) Head category	(T) Dependent position	(T) Dependent	(T) Dependent category	(T) Syntactic relation
1	gato	nome	0	o	determinante	Especificador nominal (ou determinante)
2	comer	verbo	1	gato	nome	Sujeito
4	rato	nome	3	o	determinante	Especificador nominal (ou determinante)
2	comer	verbo	4	rato	nome	Objeto direto



(T) Head position	(T) Head	(T) Head category	(T) Dependent position	(T) Dependent	(T) Dependent category	(T) Syntactic relation
1	rato	nome	0	o	determinante	Especificador nominal (ou determinante)
2	comer	verbo	1	rato	nome	Sujeito
4	comida	nome	3	o	determinante	Especificador nominal (ou determinante)
2	comer	verbo	4	comida	nome	Objeto direto
7	gato	nome	6	o	determinante	Especificador nominal (ou determinante)
4	comida	nome	7	gato	nome	Complemento nominal

PUC Minas

Similaridade Semântica

- Estruturas e suas dependências
  - **Rato** é o objeto de **comer** na primeira frase
  - **Comida** é o objeto de **comer** na segunda
- Partes do discurso
  - Substantivos, adjetivos
- Tudo isso pode ser utilizado para estimar as similares semânticas

Similaridade semântica é muito utilizada para tarefas de PLN tais como **identificação de paráfrase** e **respostas automatizadas de perguntas**

Leituras Extras

Para quem tem interesse em outros métodos de medidas de similaridade ou quer entender mais a fundo, algumas páginas:

- https://medium.com/@adriensieg/text-similarities-da019229c894
- https://code.google.com/archive/p/word2vec/