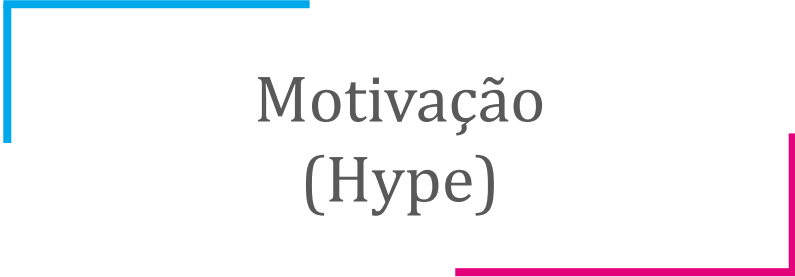


data

Processamento de Linguagem Natural (PLN/NLP)

João Pedro (Dora) Mattos • 02/06/2021

@joaopedromattos



Motivação
(Hype)





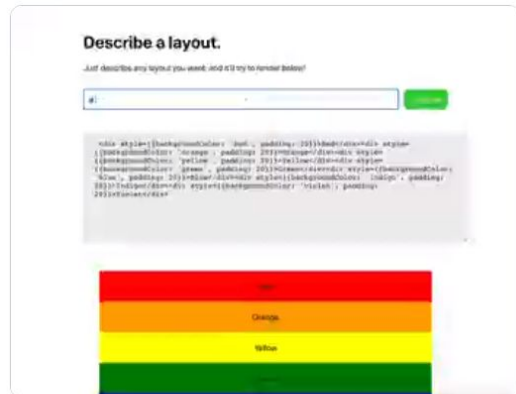
Sharif Shameem
@sharifshameem



This is mind blowing.

With GPT-3, I built a layout generator where you just describe any layout you want, and it generates the JSX code for you.

W H A T



11:01 AM · Jul 13, 2020



42.2K 691 Copy link to Tweet

GPT-3 - Geração de HTML + CSS

(Geração de texto)



AI Dungeon - Geração de histórias de RPG de Mesa

(Geração de texto)

You are Dora, The Explorer, a spy living in Chicago. You have a concealed pistol and a syringe of poison. You listen to the Russian diplomats and hear them discussing plans to assassinate the head of the German army, von Hindenburg. You decide to help the Russians fulfil their plan. You approach a group of soldiers outside a Berlin hotel.

Tip: Remember to start a "do" input with a verb, ex: Attack the orc



Do What do you do?



<div> <div>▲</div> <div>text_en</div> <div>≡</div> </div> <div>texto em inglês</div>	<div> <div>▲</div> <div>text_pt</div> <div>≡</div> </div> <div>texto em português</div>	<div> <div>▲</div> <div>sentiment</div> <div>≡</div> </div> <div>rótulo do texto, que pode ser "pos" ou "neg"</div>
<div>49043</div> <div>unique values</div>	<div>49045</div> <div>unique values</div>	<div>neg</div> <div>50%</div> <div>pos</div> <div>50%</div>
Once again Mr. Costner has dragged out a movie for far longer than necessary. Aside from the terrifi...	Mais uma vez, o Sr. Costner arrumou um filme por muito mais tempo do que o necessário. Além das terr...	neg
This is an example of why the majority of action films are the same. Generic and boring, theres real...	Este é um exemplo do motivo pelo qual a maioria dos filmes de ação são os mesmos. Genérico e chato, ...	neg

IMDB PT - Classificação de reviews de filmes

(Análise de Sentimentos)



Named Entity Recognition

In fact, the **Chinese** **NORP** market has the **three** **CARDINAL** most influential names of the retail and tech space – **Alibaba** **GPE** , **Baidu** **ORG** , and **Tencent** **PERSON** (collectively touted as **BAT** **ORG**), and is betting big in the global **AI** **GPE** in retail industry space . The **three** **CARDINAL** giants which are claimed to have a cut-throat competition with the **U.S.** **GPE** (in terms of resources and capital) are positioning themselves to become the 'future **AI** **PERSON** platforms'. The trio is also expanding in other **Asian** **NORP** countries and investing heavily in the **U.S.** **GPE** based **AI** **GPE** startups to leverage the power of **AI** **GPE** . Backed by such powerful initiatives and presence of these conglomerates, the market in APAC AI is forecast to be the fastest-growing **one** **CARDINAL** , with an anticipated **CAGR** **PERSON** of **45%** **PERCENT** over **2018 - 2024** **DATE** .

To further elaborate on the geographical trends, **North America** **LOC** has procured **more than 50%** **PERCENT** of the global share in **2017** **DATE** and has been leading the regional landscape of **AI** **GPE** in the retail market. The **U.S.** **GPE** has a significant credit in the regional trends with **over 65%** **PERCENT** of investments (including M&As, private equity, and venture capital) in artificial intelligence technology. Additionally, the region is a huge hub for startups in tandem with the presence of tech titans, such as **Google** **ORG** , **IBM** **ORG** , and **Microsoft** **ORG** .





Disclaimer



Roteiro do dia

- Aquisição de dados
- Limpeza dos datasets
- Pré-processamento
- Treinamento e Avaliação



Aquisição

Aquisição do Corpus

- Scraping e Crawling
 - Selenium, Requests e BeautifulSoup
- Datasets já conceituados academicamente
 - Bergen Corpus of London Teenage Language (COLT), IMDB Reviews, Stanford Sentiment Treebank
- Separação
 - Dados já se encontram num banco de dados. SQL / Spark



Aquisição do Corpus

- Aquisição baseada em tarefa
 - Exemplo: Geração de Texto vs. Análise de Toxicidade



Hän on journalisti. Hän on johtaja. Hän on uupunut. Hänellä on lapsenlapsi. Hän tekee töitä. Hänellä on päänsärkyä. Hänellä on hieno auto. Hän hoitaa lasta. Hän hoitaa hommat.



Kamera



Keskustelu



Litteroi



He is a journalist. He is a leader. She is exhausted. She has a grandchild. He works. She has a headache. He has a great car. She is taking care of the child. He takes care of things.



dúvidas?



Limpeza

Limpeza

- Remoção de StopWords
 - “~~As~~ rodovias selecionadas ~~para~~ receber ~~o~~ sinal foram ~~as~~ consideradas estratégicas ~~para~~ o transporte ~~de~~ passageiros ~~e~~ ~~para~~ o escoamento ~~da~~ produção agropecuária.”
- Remoção de xingamentos e palavras de baixo calão
 - [Our List of Dirty, Naughty, Obscene, and Otherwise Bad Words](#)





dúvidas?



Pré-processamento

Bag-of-Words (BOW)

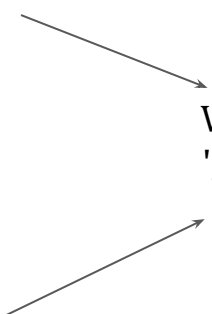
“Eu gosto muito de tomar açaí.”

“Mas também gosto muito de tomar sorvete”



Bag-of-Words (BOW)

“Eu gosto muito de tomar açaí.”



Vocabulário: {'Eu', 'Mas', 'açaí', 'de',
'gosto', 'muito', 'sorvete', 'também',
'tomar'}

“Mas também gosto muito de tomar sorvete”



Bag-of-Words (BOW)

“Eu gosto muito de tomar açaí.”

{'Eu': 1, 'Mas': 0, 'açaí': 1, 'de': 1, 'gosto': 1, 'muito': 1,
'sorvete': 0, 'também': 0, 'tomar': 1}

Vocabulário: {'Eu', 'Mas', 'açaí', 'de',
'gosto', 'muito', 'sorvete', 'também',
'tomar'}

“Mas também gosto muito de tomar sorvete”

{'Eu': 0, 'Mas': 1, 'açaí': 0, 'de': 1, 'gosto': 1, 'muito': 1,
'sorvete': 1, 'também': 1, 'tomar': 1}



Bag-of-Words (BOW)

- Vocabulário grande -> Vetores grandes -> *Curse of Dimensionality*
- Necessidade de normalizar a importância das palavras
- Não diferencia palavras comuns e palavras mais específicas
 - TF-IDF aborda exatamente esse problema





dúvidas?



TF-IDF (Term Frequency–Inverse Document Frequency)

$$tfidf(t, d, D) = \underbrace{tf(t, d)} \cdot \underbrace{idf(t, D)}$$

Quantas vezes o termo t aparece no documento d

Número de termos do documento d

$$\log_2 \left(\frac{\text{Número de documentos}}{\text{Número documentos em que } t \text{ aparece}} \right)$$



TF-IDF (Term Frequency–Inverse Document Frequency)

$$\text{Importância da palavra "Açaí" no primeiro documento} = \frac{1}{6} \cdot \log_2 \left(\frac{2}{1} \right) = 0.167$$

$$\text{Importância da palavra "Muito" no primeiro documento} = \frac{1}{6} \cdot \log_2 \left(\frac{2}{2} \right) = 0$$





dúvidas?



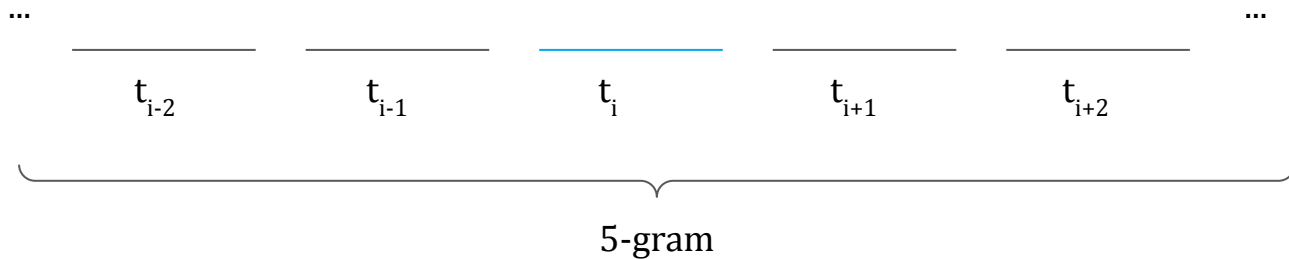
Ainda temos problemas...

- Vocabulário grande -> Vetores grandes -> *Curse of Dimensionality*
 - Representa o documento inteiro de uma vez
 - Representação literal tende a ser ineficiente



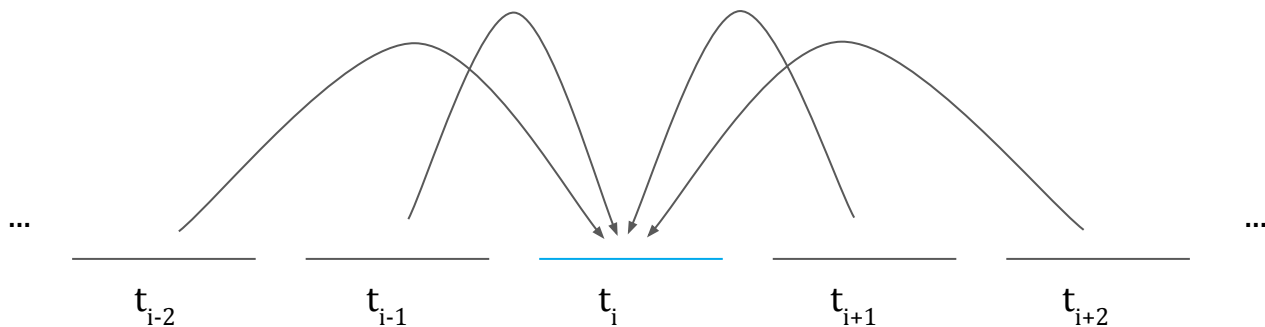
Word2Vec

- *“Diga-me com quem tu andas, e te direi quem tu és”*



Word2Vec

- *“Diga-me com quem tu andas, e te direi quem tu és”*



Word2Vec

$$L(\theta) = \prod_{i=1}^T \prod_{\substack{-m \leq j \leq m \\ j \neq 0}} P(t_{i+j}|t_i)$$

Aplicando o log, tirando a média e multiplicando por -1

$$J(\theta) = -\frac{1}{T} \sum_{t=1}^T \sum_{\substack{-m \leq j \leq m; \\ j \neq 0}} \log P(t_{i+j}|t_i)$$



Word2Vec

$$L(\theta) = \prod_{i=1}^T \prod_{\substack{-m \leq j \leq m \\ j \neq 0}} P(t_{i+j}|t_i)$$

Aplicando o log, tirando a média e multiplicando por -1

$$J(\theta) = -\frac{1}{T} \sum_{t=1}^T \sum_{\substack{-m \leq j \leq m; \\ j \neq 0}} \log P(t_{i+j}|t_i)$$

Como calcular?

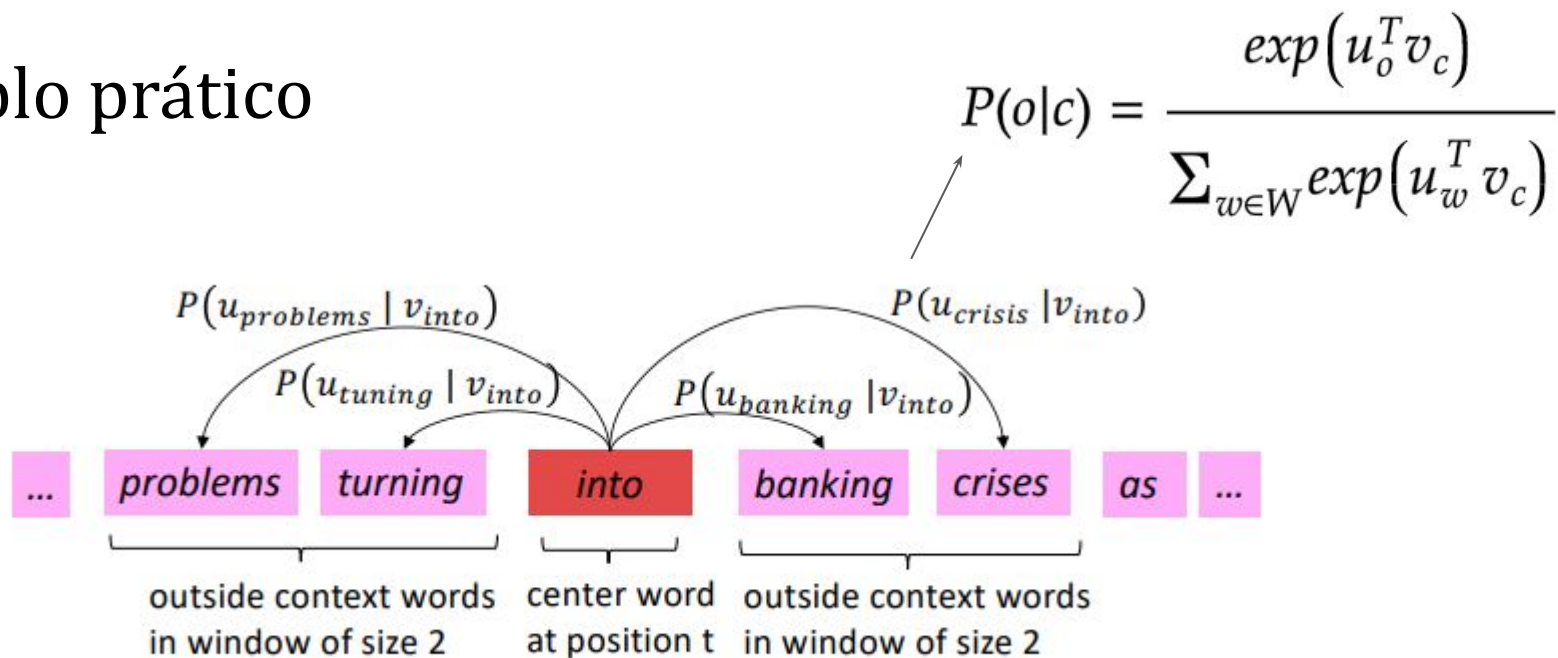


- $u \rightarrow$ vetor que representa a palavra quando está na vizinhança
- $v \rightarrow$ vetor que representa a palavra do centro

$$\text{softmax} = P(o|c) = \frac{\exp(u_o^T v_c)}{\sum_{w \in W} \exp(u_w^T v_c)}$$



Exemplo prático



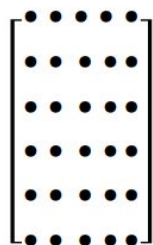
$$J(\theta) = -\frac{1}{T} \sum_{t=1}^T \sum_{\substack{-m \leq j \leq m; \\ j \neq 0}} \log P(t_{i+j} | t_i)$$



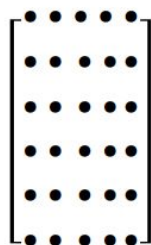
Detalhes de implementação

- Negative-sampling
- Implementação vetorizada

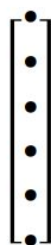
$$P(o|c) = \frac{\exp(u_o^T v_c)}{\sum_{w \in W} \exp(u_w^T v_c)}$$



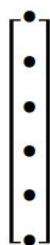
U
outside



V
center



$U \cdot v_4^T$
dot product



$\text{softmax}(U \cdot v_4^T)$
probabilities





dúvidas?



Treinamento e Avaliação

Treinamento

- Muitas tasks diferentes dificultam a escolha do modelo
- Como os modelos que você está considerando se adaptam à task?
 - Vale a pena consultar na literatura.
- **Faça baselines!**
 - Canivetes suíços de PLN para Python: NLTK, Gensim e Spacy



Avaliação

- Temos muitas tasks diferentes -> Muitas métricas diferentes
- Procurar na literatura é fundamental
 - BLEU para tradução, GLUE métrica geral de entendimento de linguagem, etc..



Referências

- [Stanford CS224N - Lecture 1](#)
- [Stanford CS224N - Lecture 2](#)
- [The Illustrated Word2Vec - Jay Alammar](#)

