Insper - Engenharia de Computação - 9° Semestre

# Projeto 3
# Sumarização

Bruna Kimura e Elisa Malzoni

NLP - Prof. Fábio Ayres

# Sumário

**Introdução** **2**

**Metodologia** **2**

Sumarização 2

Clustering 2

TF-IDF 3

CBOW 3

TextRank 3

Extração de Palavras-Chave 3

**Resultados** **4**

Sumarização 4

Documento 1 4

Documento 2 6

Extração de palavras-chave 7

**Conclusão** **8**

# Introdução

Nesse projeto foram realizadas a sumarização de texto e a extração de palavras-chave.

Existem dois tipos de sumarização de texto a extrativa e a abstrativa. No primeiro modelo, extrativa, as frases que compõem o sumário estão presentes no próprio texto do documento, ou seja, são extraídas do texto fonte. Já na segunda técnica, abstrativa, os textos que compõem o sumário não necessariamente estão inseridas no texto. Nesse tipo de sumário é necessário criar um texto novo que consiga abstrair o conteúdo do texto fonte.

A sumarização feita neste trabalho foi a extrativa, para tanto, utilizou-se duas metodologia diferentes, o *clustering* e o *TextRank*.

Por fim, a última etapa foi a extração de palavras-chave. O objetivo é bastante semelhante à sumarização, mas ao invés de buscar por sentenças que resuma o texto, a intenção é retirar algumas palavras que possam indicar qual o assunto do documento.

Dessa forma, este trabalho é um experimento que tem como objetivo entender como cada aplicação e metodologia se comportam em determinados conjuntos de texto e quais as vantagens e desvantagens de se utilizar cada uma dessas abordagens.

# Metodologia

O projeto foi dividido em duas partes. A primeira diz respeito a sumarização de documentos, enquanto a segunda parte está relacionado a extração de palavras-chave de documentos.

O *dataset* utilizado foi o CNN/Daily Mail, esse documento traz notícias do jornal CNN. Utilizou-se 100 documentos da seção *Stories*, que estão disponíveis no repositório do grupo. Desses documentos pegou-se somente o corpo da notícia, sem os *highlights*.

Para analisar o desempenho, será feita uma análise qualitativa dos sumários e das palavras-chave extraídas. A ideia é comparar para certos documentos quais sumários são mais adequados para resumir o assunto principal do texto.

## Sumarização

A sumarização do documento foi feita utilizando duas metodologias diferentes (*Clustering* e *TextRank*), a fim de identificar a eficácia de cada uma das abordagens.

### *Clustering*

A primeira metodologia para a sumarização de documentos é a utilização da combinação entre vetorização e clusterização. A ideia dessa abordagem é criar alguns *clusters* para separar as sentenças de um documento em grupos de assunto. Então, retira-se a sentença mais relevante de cada grupo, ou seja, a que está mais próxima do centro do *cluster,* e

utiliza-a como sumário do documento. Foi utilizado dois métodos distintos para a vetorização das sentenças.

### TF-IDF

Primeiramente, foi utilizado o método de vetorização TF-IDF (*term frequency–inverse document frequency*). Nesse tipo de abordagem o primeiro passo foi criar os vetores de sentenças para cada documento utilizando a vetorização TF-IDF. Com esse vetor feito, a próxima etapa foi então encontrar os possíveis *clusters* para cada um desses textos. Foi decidido que o total de *clusters* por documento seriam dois, uma vez que os textos não são muito longos, portanto, duas sentenças foram o suficiente para conseguir sumarizá-lo. Com os *clusters* já montados, escolheu-se a sentença que mais se aproximava do seu centro, totalizando duas sentenças para o sumário.

### CBOW

Já a segunda metodologia utilizada foi o CBOW (*Continuous Bag of Words*). Nesse modelo o vetor é criado a partir da somatória de todos os *embedding* de cada palavra de todos os documentos.

Com o vetor de CBOW feito, o próximo passo é semelhante com o feito no TF-IDF. Da mesma maneira, é necessário criar os *clusters* para cada um dos documentos, nesse caso, também foi escolhido apenas dois *clusters*, já que os textos são os mesmos que os utilizados no método acima. Para cada *cluster*, buscou-se as sentenças que mais se aproximavam do centróide do *cluster*. Dessa forma, cada documento possui duas sentenças que sumarizam o texto.

### *TextRank*

Para fazer a sumarização com o TextRank foi utilizado esse tutorial. Contudo no caso do tutorial o sumário era de todos os documentos e no do caso do projeto foi feito um sumário por documento.

Nesse algoritmo é feita a similaridade entre sentenças, utilizando uma matriz de similaridade, que é composta pelas semelhanças de cosseno dos vetores das sentenças. Essa matriz é então transformada num grafo, neste grafo utiliza-se o algoritmo de PageRank para ranquear as sentenças. Depois de ordenadas, as três primeiras sentenças formam o sumário.

## Extração de Palavras-Chave

Para a extração de palavras-chave foi utilizado uma metodologia muito semelhante a de sumarização. Foram adaptadas duas das metodologias de sumarização para esse experimento, o TF-IDF e *TextRank*.

No primeiro método foi feito o *clustering* utilizando a vetorização do TF-IDF. Diferentemente da sumarização, nesse caso, o vetor é composto de palavras e foram

criados 10 *clusters*. O restante do algoritmo segue a mesma lógica que a da sumarização utilizando TF-IDF.

Já o *TextRank*, ao invés de usar os vetores de sentenças como na sumarização utiliza-se os vetores de palavras. Então a única diferença com o algoritmo de sumarização é na matriz de similaridade que possui as semelhanças de cosseno dos vetores das palavras. Essa matriz também foi transformada em grafo, ranqueou-se as palavras e as 10 primeiras palavras são as palavras-chave do documento.

# Resultados

## Sumarização

A fim de comparar a qualidade dos sumários gerados por cada uma das aplicações foram escolhidos dois documentos que representasse o comportamento geral dos resultados.

A intenção era entender as vantagens e desvantagem de cada abordagem e verificar qual se adequa melhor ao *dataset* teste.

### Documento 1

arquivo: `0a00a9aebcb754c51534867cf1db2335dcb76884.story`

"Manchester United's Spaniards may have struggled to gel on the pitch, but they have been building up a close relationship off it.", 'Juan Mata, Ander Herrera and David de Gea have become good friends and enjoyed lunch together in Hale Village, Cheshire on Tuesday afternoon.', "The Spanish trio were all named in Louis van Gaal's starting line-up for the opening day fixture against Swansea, although they were unable to prevent a disappointing 2-1 defeat bySwansea.", 'VIDEO Scroll down to watch Mata challenge compatriot De Gea to Corner Kick Challenge\xa0', 'Meet and greet: Juan Mata (left) has been helping fellow Spaniard Ander Herrera (right) settle in', 'Join the club: Manchester United keeper David de Gea and Herrera pictured together on Tuesday', 'Three Amigos: The Spaniards stroll down the high street as they enjoy an afternoon in Hale Village', "Van Gaal has brought another Spanish speaker to Old Trafford, with Argentina defender Marcos Rojo describing it as a 'dream' to play for the Old Trafford club.", 'The Sporting Lisbon player has emerged has been a key summer target for Van Gaal and United have paid £16million for the defender and sent Portuguese winger Nani back to his former club on a one-year loan deal.', "In an interview with the radio station Continental, quoted in several national newspapers, Rojo said: 'It's a dream to play at Manchester United and I am very proud of having the chance of working with [Louis] van Gaal.", '"I spoke with Juan Sebastian Veron about Manchester United when we were at Estudiantes. I have always liked English football, and I should adapt to this new playing style easily.'\xa0", 'Double act: Mata and Herrera are hoping to form a close-knit midfield partnership at Old Trafford', 'Slow start: Mata and his United team-mates will be hoping for  improved performances after defeat by Swansea', 'Dream move: Marcos Rojo (left) has signed for United from Sporting Lisbon for £16m', 'VIDEO United reach deal for Rojo'\xa0'"

A sumarização feita pelo:

- Clustering (TF-IDF)

> “The Sporting Lisbon player has emerged has been a key summer target for Van Gaal and United have paid £16million for the defender and sent Portuguese winger Nani back to his former club on a one-year loan deal.
>
> VIDEO United reach deal for Rojo'\xa0”’

- Clustering (CBOW)

> “The Sporting Lisbon player has emerged has been a key summer target for Van Gaal and United have paid £16million for the defender and sent Portuguese winger Nani back to his former club on a one-year loan deal.
>
> I spoke with Juan Sebastian Veron about Manchester United when we were at Estudiantes. I have always liked English football, and I should adapt to this new playing style easily.”

- TextRank

> “The Spanish trio were all named in Louis van Gaal's starting line-up for the opening day fixture against Swansea, although they were unable to prevent a disappointing 2-1 defeat bySwansea.
>
> Van Gaal has brought another Spanish speaker to Old Trafford, with Argentina defender Marcos Rojo describing it as a 'dream' to play for the Old Trafford club.
>
> The Sporting Lisbon player has emerged has been a key summer target for Van Gaal and United have paid £16million for the defender and sent Portuguese winger Nani back to his former club on a one-year loan deal.”

De forma geral é possível notar que pelo menos uma sentença se repete em todas as abordagens, essa sentença é bastante relevante no texto e mostra que todos os três métodos possuem alguma eficiência em sumarização.

Na primeira abordagem (TF-IDF) é possível notar que apesar da primeira sentença ser bastante relevante a segunda sentença não é algo que deveria estar no sumário, já que é apenas uma indicação para um vídeo. Enquanto na segunda abordagem (CBOW) a sumarização se mostrou mais eficiente. A primeira sentença é a que se repete, porém a segunda sentença é mais relevante que no caso anterior, apesar de não ser um bom trecho para resumir a notícia. Já no TextRank, as sentenças escolhidas foram bem mais pertinentes. Fora a sentença que se repete, há sentenças que mostram sobre quem a notícia se referência e suas ações, conseguindo assim, transmitir bem a ideia do noticiário.

Portanto, neste primeiro caso é possível verificar que a abordagem TF-IDF é a menos eficiente. A CBOW apesar de ser a melhor no método de *clustering* ainda assim se mostrou menos eficaz em comparação a *textRank*, que resultou em sumário bastante interessante e relevante com a notícia.

## Documento 2

arquivo: `0a0b2b48245b8e585fed35eb89f69a8b1f0ee25a.story`

'Mixed martial arts fighter Anderson Silva says he will fight for a spot in the Brazilian taekwondo team at the 2016 Olympics in Rio de Janeiro.', 'The announcement was made on Wednesday after a meeting with Brazilian taekwondo officials.', "Considered one of the best pound-for-pound fighters in the history of mixed martial arts, Silva said he is 'trying to give back to the sport' in which he began his career.", 'Anderson Silva met with Brazilian taekwondo officials and will compete for a spot in the\xa02016 Olympics team', "The former UFC champion said he is 'trying to give back to the sport' in which he began his career", "Taekwondo confederation president Carlos Fernandes said it will be an 'honor for our sport to welcome an athlete of this importance.'", "However, he also made it clear that Silva will have to fight his way into the Olympics and won't be helped because of his UFC stardom.", 'Silva is a taekwondo ambassador and a black belt in the sport.\xa0', 'The former UFC champion tested positive for two steroids in an out-of-competition test Jan. 9, and also failed a test after his UFC victory over Nick Diaz on Jan. 31.\xa0', 'The 40-year-old Brazilian posted a photo of himself via his Twitter page practicing taekwondo last week'

A sumarização feita pelo:

- Clustering (TF-IDF)

> “Mixed martial arts fighter Anderson Silva says he will fight for a spot in the Brazilian taekwondo team at the 2016 Olympics in Rio de Janeiro.
>
> Considered one of the best pound-for-pound fighters in the history of mixed martial arts, Silva said he is 'trying to give back to the sport' in which he began his career.”

- Clustering (CBOW)

> “Taekwondo confederation president Carlos Fernandes said it will be an 'honor for our sport to welcome an athlete of this importance.
>
> The 40-year-old Brazilian posted a photo of himself via his Twitter page practicing taekwondo last week”

- TextRank

> “Anderson Silva met with Brazilian taekwondo officials and will compete for a spot in the 2016 Olympics team
>
> The former UFC champion said he is 'trying to give back to the sport' in which he began his career
>
> However, he also made it clear that Silva will have to fight his way into the Olympics and won't be helped because of his UFC stardom.”

Neste caso é possível notar que não houveram sentenças semelhantes em nenhuma metodologia, apesar de todas de forma geral conseguirem destacar alguns pontos interessantes. No TF-IDF, é bastante pertinente destacar quais foram as duas sentenças escolhidas. É possível notar, nesse caso, que os vetores escolhidos possuem bastante

palavras que se repetem com frequência no texto. Essa é uma das característica dessa abordagem para selecionar os *clusters*. Nesse documento em específico essa abordagem pareceu bastante eficiente, já que as sentenças escolhidas conseguem de certa forma transmitir a ideia central do texto. Com o método CBOW, o resultado não ficou muito interessante. É possível notar que o CBOW criou dois *clusters* que possuem assuntos completamente diferentes, o que é importante para conseguir encontrar todas as ideias chaves do texto, porém nesse caso em específico, o texto estava muito focado em um único assunto. Assim, esse método não pareceu muito interessante para esse documento, já que selecionou sentenças que não transmitem os pontos centrais do documento. Por fim, mais uma vez, o que se mostrou o melhor método foi o *TextRank*. As sentenças escolhidas por essa metodologia são as mais coerentes com o documento. É possível ver que as duas primeiras sentenças transmitem as mesmas ideias que as sentenças do TF-IDF mas com trechos que fazem mais sentido avulsos. Já a última sentença tenta pegar alguma outra ideia do texto, mas continua sendo bastante relevante, diferentemente do CBOW que ao tentar buscar um outro ponto acaba pegando um trecho que não transmite um ponto importante da notícia.

## Extração de palavras-chave

Neste experimento foram utilizados dois métodos de extração de palavras-chave, o TF-IDF e *TextRank*, já que foram os métodos que mostraram comportamentos opostos na questão de eficiência de sumarização. A ideia desse teste era entender se esses métodos teriam os mesmos resultados que os apresentados na sumarização ao serem aplicados na extração de palavras-chave. Dessa forma, foi analisada a qualidade que as palavras encontradas possuíam em cada uma das extrações.

Foi retirado um exemplo de documento e as respectivas palavras-chaves encontradas para esse texto.

`doc: 0a1a94f06809b73d31cf1f43435827cd21467d94.story`

'Ryan Gorman', '15:56 EST, 18 August 2013', '| ', '04:14 EST, 19 August 2013', 'A Philadelphia woman is accused of tying her disabled cousin to a urine-soaked bed in a scheme to steal her assistance checks.', 'Regina Bennett, 46, was first arrested early Saturday morning after an altercation with her neighbour. Cops doing a walk-through of her home found a malnourished, disabled woman bound by her arms and legs with rags to a urine-soaked bed wearing only a diaper. As the handicapped woman's caregiver, Bennett had control over her cousin's finances and received her social security checks, according to reports.', 'Instead of public assistance, Bennett was handed on Sunday a litany of charges ranging from assault, to making terroristic threats, kidnapping, public drunkenness and false imprisonment.', 'Horrific discovery: Regina Bennett is in custody for both the fight and imprisoning her disabled cousin', 'Police were called just before midnight Friday night after Bennett allegedly got into a drunken brawl with a neighbor, according to WPVI. They were told by neighbors there may be a small child inside.', "She started snapping on neighbors, using profanity and names," Richard Master told Fox 29.', 'Cops searching the house of horrors said they instead found a 36-year-old woman with cerebral palsy, severe malnutrition, bed sores and other signs of abuse being kept prisoner in squalid conditions wearing only a diaper, according to NBC Philadelphia.', 'Officers untied the allegedly tortured woman and sent her to a local hospital.', "She looked like she was distraught, maybe, out of it – I guess because there was a lot of ambulances and police around her," a neighbor told Fox 29.', "Bennett was appointed caretaker of her cousin in 2009, according to NBC Philadelphia. Her role saw

her overseeing the woman's finances.", 'A relative of Bennett's defended her, saying she never saw any evidence of abuse.', "'She took care of her like she was her own flesh and blood child. Her room was neat and clean, she went to her programs and fed her - that's all I know,' the woman identified only as Phyllis told local media.", 'House of horrors: Second from left, the house where the handicapped woman was abused by Bennett', "Terrible: it hasn't yet been determined how long the relative was kept prisoner by Bennett", 'Neighbors said they barely knew Bennett.', "It was scary, it was weird, I couldn't believe it,' Keisha Gonzalez told Fox 29, adding 'you never know who you live next door to.", 'Master told WPVI that he saw 'the para-transit van drop her off but that was it.", "Bennett is currently in custody, her victim has not been publicly identified. There was also no word on who would be the disabled woman's caretaker in Bennett's absence."

- TF-IDF
  - woman
  - cousin
  - saturday
  - bennett
  - instead
  - neighbors
  - local
  - appointed
  - relative
  - flesh
- *TextRank*
  - instead
  - instead
  - took
  - would
  - yet
  - saying
  - never
  - never
  - saw
  - saw

Nesse experimento, ao contrário da sumarização, o TF-IDF obteve um resultado melhor comparado com o *TextRank*. As palavras-chave obtidas com o *TextRank* são genéricas, ou seja, elas não trazem significado desacompanhadas de outras palavras, e também mostrou algumas palavra mais de uma vez. Já no TF-IDF as palavras-chave são mais específicas, mas mesmo assim, essas palavras não refletem o conteúdo do texto.

# Conclusão

Existem várias formas de se fazer uma sumarização. Com os resultados já apresentados, e considerando o *dataset* utilizado, pode-se perceber uma tendência ao observar o desempenho dos métodos utilizados. O método de *clustering* utilizando a vetorização TF-IDF foi a que se mostrou a menos eficiente. Nesse modelo, as sentenças são escolhidas com base na frequência das palavras, que para esse *dataset* não foi uma abordagem adequada na maioria dos documentos. O método de *clustering* utilizando o CBOW teve uma eficácia melhor entre os métodos de *clustering*, porém em alguns documentos não se comportou da maneira desejada. Por fim, o *TextRank* foi o que apresentou o sumário mais adequado para esse *dataset*, os seus resultados mostraram ser eficientes em quase todos os documentos.

Já na extração de palavras-chave nenhum método se mostrou muito eficaz. Em ambos os casos, TF-IDF e TextRank, as palavras obtidas não refletiam as ideias do texto, apesar de que o primeiro método se mostrou ligeiramente superior ao segundo.