# Técnicas de aprendizado não supervisionado: Health Tweets Dataset

Farzin Shams
farzinshams95@gmail.com

## Introdução

Nesta tarefa, temos como objetivo investigar técnicas de aprendizado não supervisionado em um dataset de tweets sobre saúde que contém três arquivos: um com os tweets em forma de string, um com os tweets codificados em *bag of words*, e um com os tweets codificados em *word2vec*.

A primeira técnica investigada, kmeans, tem como objetivo encontrar grupos presentes nos dados. Ela encontra um conjunto de centróides, que são pontos que indicam a qual grupo algum outro dado em específico pertence, baseado na distância deste dado ao centróide mais próximo. Ou seja, um dado pertencerá ao grupo definido pelo centróide mais próximo.

O modelo de kmeans que deicidimos utilizar é o kmeans++. Nele, a inicialização é feita de forma mais eficiente: um dado em aleatório é escolhido para ser o primeiro centro. Em seguida, cada centro é adicionado numa posição com probabilidade proporcional ao quadrado da distância do centro mais próximo. Esta técnica melhora a convergência do algoritmo pois ela espalha os centros no espaço de dados, facilitando a convergência destes para seus respectivos grupos.

O segundo, PCA, tem como objetivo reduzir a dimensionalidade dos dados através da projeção destes em hiperplanos de dimensão menor, removendo as dimensões em que os dados têm menor variância.

## Kmeans

O conjunto de dados escolhidos para investigar o kmeans é o de codificação *word2vec*, pois ele codifica os tweets de forma sucinta e eficiente, reduzindo a variância presente nos dados e a dimensão deles, o que facilita a análise e reduz o custo computacional.

Como o kmeans++ é muito sensível a outliers, decidimos remover os dados que apresentavam algum *feature* com valor z, em módulo, maior que 4. Isto resultou na diminuição do número de dados em aproximadamente 3,3%.

A figura 1 mostra a curva da soma dos erros quadráticos em função do número de centros. Para escolher o número de clusters, usamos o método do cotovelo: selecionamos o valor a partir do qual o incremento do número de clusters tem influência mínima em reduzir o SSE. O valor escolhido, então, foi igual a 29.

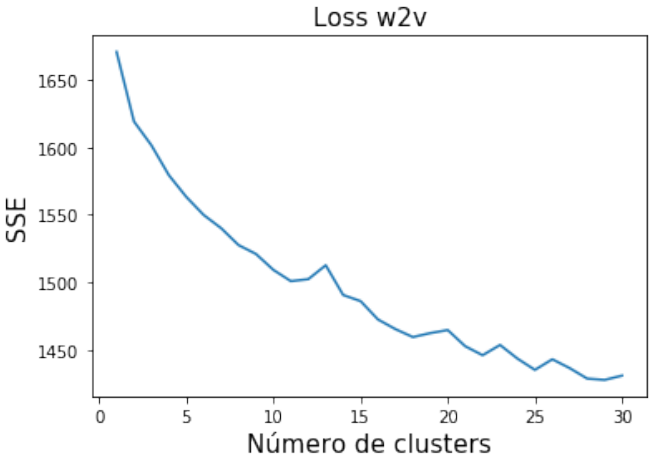


Figura 1. SSE em função do número de centróides

Com os clusters à disposição, escolhemos três deles para analisar os tweets a eles pertencentes a fim de verificar se eles são parecidos. A seguir, listaremos três tweets para cada cada grupo escolhido de forma aleatória.

- **Cluster 1**
- @RoscoeTheHorse Not really. But common sense says to avoid anyone who's sick. Virus spreads only when there are symptoms! EbolaQandA
- 1 in 8 U.S. babies is born pre-term. @LIFE takes a look at how hospitals saved these tiny humans 75 yrs ago
- RT @CNNVideo: Deaf toddler's reaction to hearing his dad's voice for the first time will make your day. @drsanjaygupta reports:
- **Cluster 4**
- RT @kellywallacetv: What's the advice when your child wants to stop a physical activity amp; you know it's good for them to keep doing it? fi…
- @jillianmichaels shows us 3 simple moves to workout your chest. These can be done from anywhere in a few minutes:
- RT @drsanjaygupta: letting @diananyad get some rest, and then I sit down with her for 1st post swim intvu. anything you want me to ask? ex…
- **Cluster 17**
- RT @NCADASTL: @cnnhealth Thanks for the article "What you need to know about synthetic drugs". Great info for parents amp; kids.
- FDA warned docs this week about prescribing too much

acetaminophen. 5 things you need to know about this pain killer
- Worried about the new virus sweeping the Midwest? Here's what you need to know to keep your kids (amp; you) healthy:

Aparentemente, existe uma semelhança no assunto entre os tweets de cada cluster. No cluster 1, o segundo e terceiro tweet falam a respeito da saúde de bebês e crianças. No cluster 4, o tema central aparenta ser sobre atividade física e descanso. No cluster 17, os dois primeiros tweets comentam sobre remédios.

Como forma objetiva de medir a qualidade dos clusters, iremos usar o coeficiente Silhouette. Este coeficiente é definido no intervalo [-1,1], onde uma valor igual a 1 indica um cluster com dados bem semelhantes entre si e diferentes dos dados pertencentes a outros clusters.

Para cada cluster, usamos o dado mais próximo ao centróide para calcular este coeficiente. Assim, obtemos 29 valores destes, e a média foi igual a $0.0569$, que indica que o algoritmo de clusterização tem capacidade mediana de separar os dados em clusters.

## PCA

Para investigar a influência da redução de dimensionalidade dos dados, usamos o método de PCA. Seguimos a mesma série de procedimentos feitos anteriormente em dois conjuntos de dados reduzidos: um com 80% da variância original, e um com 95%. O primeiro diminuiu o número de features originais, 128, para 84; e o segundo diminuiu para 115

As Fig. 2 e 3 mostra o SSE em função do número de clusters para o PCA com 80% e 95% da variância total, respectivamente. Como podemos perceber, a curva é parecida com o caso com todos os features e, assim, o número de clusters selecionado foi igual ao caso anterior: 29.

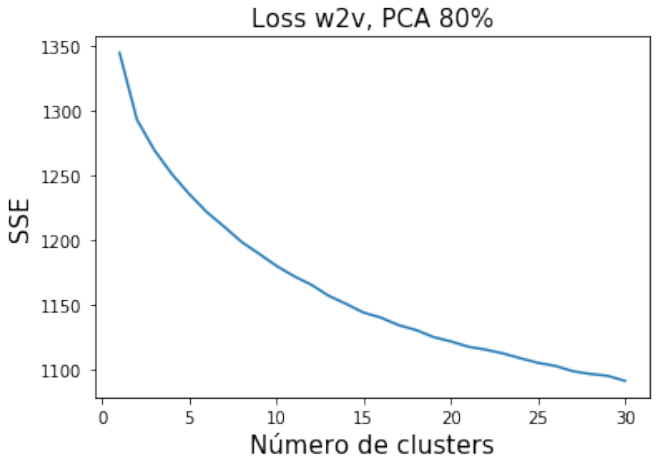


Figura 2. SSE em função do número de centróides para PCA com 80% da variância

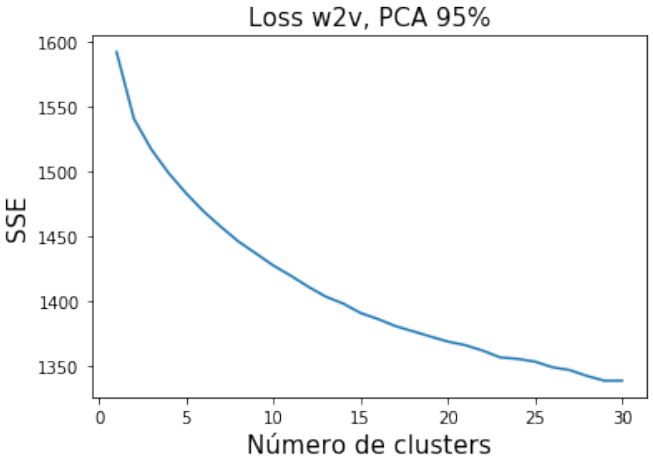


Figura 3. SSE em função do número de centróides para PCA com 95% da variância

A seguir, listeramos três tweets de três grupos para o PCA 80%:

- **Cluster 1**
- Gun violence in PG-13 movies has more than tripled since 1985. Do you let your kids watch violent films?
- More than 1 in 3 U.S. adults have prediabetes, and the numbers are rising. Learn how to eat to lower your risk:
- Craving chocolate? Take a walk! Data finds that people who take walks snack less under stress. More tips:
- **Cluster 4**
- Every gym has them! The 8 most annoying people at the gym, and how to deal with your frustrations:
- RT @Robinsbite: A7: And, beware the chip/bread/cracker basket. Here in Texas, chips can be the death of people when dining out! healthtalk
- RT @ChristysChomp: healthtalk A7: I say, save the calories for the food! Don't waste them on sugary drinks (sodas, sweet teas, creamy/sweet coffee, etc).
- **Cluster 17**
- What's it like inside the Ebola hotzone? An American doctor who had Ebola is back in Africa
- .@WHO says Ebola outbreak in West Africa is one of the "most challenging"they've ever faced
- WHO Ebola response chief says virus still spreading due to lack of change in behaviors

Analisando os tweets, percebemos que só no cluster 17 há um claro tema central: o surto do ebola. Usando o mesmo procedimento do cálculo do valor médio do coeficiente de Silhouette do caso anterior, obtemos um valor igual a $0.04984$, um valor um pouco menor que o anterior. Continuando, a seguir, listaremos os tweets do conjunto PCA 95%:

- **Cluster 1**
- IBD? What food do you miss most? You could be eating it after a recipe makeover by @joybauer. @ us your dish and joyrecipemakeover
- RT @DavidKirsch: @cnnhealth they know that movement is good for their body - heart, brain and muscles.

and they like doing it too! fitfam…
- RT @CSPI: A7 So to watch calories while dining out, skip the apps and desserts, and split entrees or take some home. healthtalk
- **Cluster 4**
- .@WHO says Ebola outbreak in West Africa is one of the "most challenging"they've ever faced
- What's it like inside the Ebola hotzone? An American doctor who had Ebola is back in Africa
- .@HPutt @CDCgov warns against any nonessential travel to the region but has not banned outright. Quarantine period is 3 wks for Ebola
- **Cluster 17**
- When you're sick, the last thing you want to do is get out of bed to see a doctor. These apps bring the doc to you!
- RT @jdwilson2: Do you have a favorite T-shirt you just can't give up? Share a pic amp; what it says about you using TshirtTales
- Something on your mind? Know of a topic you'd like to see covered? Now you can e-mail us at OPED.Health@cnn.com. We'd love to hear from you!

Analisando os tweets, percebe-se que o tema central do cluster 1 é sobre comida e dieta; no cluster 4, é sobre o surto de ebola. O coeficiente de Silhouette médio para este conjunto é igual a 0.025577.

Como podemos perceber, o uso do PCA não possibilitou melhora significativa nos resultados obtidos. Uma possível explicação para isso é que os features com maior variância não necessariamente possuem a informação mais importante para a separação dos dados.

## CONCLUSÃO

O kmeans usado neste dataset teve desempenho regular, mostrando algum sinal de vida inteligente na separação de dados, julgando pelo tema central do alguns tweets do mesmo cluster. Uma possível explicação, então, é que este dataset não apresenta um conjunto bem separável de dados, julgando também pela falta de 'cotovelo' no gráfico do SSE em função do número de clusters.

O PCA, também, não melhorou a capacidade de separação dos dados de forma significativa. Porém, ela agilizou a convergência dos resultados por diminuir o número de features.