Insper 2020

Relatório de Processamento de Linguagem Natural: Técnicas de sumarização

Grupo: Hugo Mendes, Jean Walper
Disciplina: Processamento de Linguagem Natural
Professor: Fábio Ayres
Data: 14/05/2020

Introdução

Um algoritmo de sumarização de texto é capaz de analisar as informações contidas em um texto e extrair dela um parágrafo curto, preciso e sucinto que contenha o maior número possível de informações relevantes ao leitor.

Textos sumarizados estão presentes no nosso cotidiano. Nem sempre iremos ler um texto muito longo, por questão de tempo ou até de interesse, então um texto compacto e apenas com a informação relevante é útil para não perder a atenção do leitor antes de passar a informação, por exemplo em notícias e manchetes ou sinopses de filmes.

Outro exemplo de sumarização seriam anotações, como as anotações de um estudante, onde apenas as informações são anotadas mas geralmente com menos ênfase em coesão ou gramática. Esse comportamento é útil para uma pessoa que já tem familiaridade com o assunto resumido, porém talvez não seja tão útil para pessoas que não estavam presentes na aula.

Nos final dos anos 1950, começaram a surgir estudos sobre como computacionalmente realizar uma sumarização de textos automatizada. Estudos como o artigo de Hans Peter Luhn, chamado “The automatic creation of literature abstracts” iniciaram uma busca por métodos precisos de sumarização.

Essas técnicas de sumarização podem ser divididas primordialmente entre dois grupos, Sumarização Extrativa e Sumarização Abstrativa.

A principal diferença entre esses dois grupos é que em uma sumarização Extrativa, o resumo gerado contém apenas sentenças tiradas do próprio texto, destacando as mais importantes. Uma sumarização Abstrativa por outro lado, é capaz de gerar sentenças novas para passar a informação ao leitor.

Nesse projeto, o grupo focou em utilizar dois algoritmos diferentes, ambos capazes de realizar uma abstração Extrativa. O “Clustering” e o “Textrank”.

O Clustering funciona analisando as sentenças vetorizadas, e utilizando distâncias dentro desse espaço vetorial para detectar “clusters”, ou agrupamentos de pontos nesse espaço.

Cada ponto é uma sentença. E a proximidade entre eles revela similaridades entre os assuntos abordados. É definido previamente o número N de clusters desejado, então o algoritmo separa os pontos em N grupos, para cada um dos grupos, é escolhido o ponto mais central, referenciando uma sentença que representa aquele grupo. Então as sentenças escolhidas são apresentadas na mesma ordem que aparecem no texto original.

A lógica por trás do algoritmo é retirar redundância e tentar utilizar uma sentença para cada tópico abordado do texto, pegando N sentenças distantes mas com alta representatividade.

O Textrank ganhou esse nome por ser baseado em um conhecido algoritmo chamado Pagerank, utilizado principalmente para ranquear páginas web em mecanismos de busca. Então para entender o Textrank é importante entender antes o Pagerank:

De forma simplificada, o Pagerank avalia uma página web pela probabilidade de um usuário navegando na internet chegar nessa página através de links contidos em outras páginas. Uma matriz é construída contendo, para cada página, a probabilidade de um usuário partir da página atual para cada uma das outras páginas contidas nos eixos da matriz.

Para fins de simplicidade, imagine que a probabilidade de um usuário clicar em um desses links é de 100% e é perfeitamente distribuída entre todos os links contidos na página. Dessa forma, as somas dos valores de um eixo dessa matriz sempre será 1, já que um eixo contém sempre a soma de partir de uma página para qualquer outra página considerada no algoritmo. Agora a soma do outro eixo é o valor utilizado para o ranqueamento, já que um valor alto nessa soma indica que a página é mencionada em várias outras páginas via links, e portanto tende a ser uma página mais importante e visitada.

A ideia então do Textrank é fazer uma matriz parecida com essa utilizada no Pagerank, porém ao invés dos eixos serem páginas web, cada eixo é uma sentença, e o valor que cada célula representa é a similaridade entre as duas sentenças vetorizadas (Cosine Similarity).

O Textrank então usa a soma de um dos eixos da matriz (tanto faz o eixo já que a matriz será simétrica em relação à diagonal principal.) para dizer quais são as sentenças mais importantes e coloca elas no resumo.

Métodos

O grupo optou por tentar sumarizar um dataset de análises de jogos digitais. Se colocando na posição de um desenvolvedor que publicou um jogo e está tentando otimizar o tempo necessário para receber o maior feedback possível lendo tais análises uma a uma.

Esse desenvolvedor então usaria o algoritmo estudado neste projeto, e o tempo necessário para ler todas as análises iria diminuir consideravelmente, mas a quantidade de informação contida nos textos sumarizados não seria muito inferior à quantidade de informação contida em todas as análises.

Utilizamos um dataset de análises australianas da plataforma de distribuição de jogos digitais para computador, a Steam. Nos primeiros testes percebemos que a característica de uma review de jogos é de ter textos curtos, o que não faz sentido para quando se está testando um algoritmo “encurtador” de texto. Para resolver o problema aplicamos um threshold mínimo de 15 sentenças, para que a review entrasse na lista de documentos válidos para teste. Este corte fez com que passassemos de um total de 59305 para 766 documentos elegíveis a sumarização. O link do dataset utilizado pode ser encontrado em:

http://deepx.ucsd.edu/public/jmcauley/steam/australian_user_reviews.json.gz

O objetivo é que a sumarização seja capaz de conter a opinião geral do autor da análise, e ainda se possível, apontar quais motivos levaram a essa opinião. Um resumo com essas características é considerado satisfatório.

Para validar a eficácia dos algoritmos implementados, optamos por fazer uma análise qualitativa das sumarização geradas. A definição de eficácia é empírica e subjetiva. Adotou-se como método, a leitura do texto original, a leitura da sumarização via clustering e por fim a leitura da sumarização via TextRank. Com a leitura dos 3 elementos, discutimos qual aparentou ter a maior quantidade de informação e ainda mantendo minimamente a coesão, e levando em consideração as características que tornam o resumo satisfatório.

O link para o repositório git do projeto, pode ser encontrado em:

https://github.com/hugosoftdev/estudo-sumariza-o

Resultados

Por meio de vários testes, pudemos identificar que os melhores resultados foram obtidos com a junção das 3 frases principais destacadas por ambos os algoritmos. Para análise dos mesmos, segue uma amostra dos textos sumarizados:

Review 1:

"Buy it. Support it. It's awesome. I will find you if you don't.^ Kek, what an old report.Personally I think you shouldn't play this game, although it might get better in a few years.I bought this game as soon as it came out, I bought the founders two pack. It was fun, pretty buggy, but it was full of promise. Skip forward a few years later. The game is worse. Lets talk about starting the game. The game is slow to start. Starting a game will take just as long, and when you get into the game itself, prepare for unoptimized slowness. The game runs terribly, is riddled with bugs and the devs generally don't add any new content. If this was developed by a good indie team, or an AAA team (Oh, did i forgot to mention they fired several people who worked at Bioware, and hired some scrubs?), it would be amazing. The concept is great, I would have really looked forward to this game shining, but the devs are horrible. I do not recommend this game."

Sumarização baseada em clustering:

"The concept is great, I would have really looked forward to this game shining, but the devs are horrible.The game is slow to start.It's awesome."

Sumarização baseada em TextRank:

"The game is slow to start.Starting a game will take just as long, and when you get into the game itself, prepare for unoptimized slowness.The concept is great, I would have really looked forward to this game shining, but the devs are horrible."

Review 2:

"Not the best shooter, nor will you have the "funniest" time you can get out of an FPS. If you're looking to have fun with mate or just by yourself, play, BF4, COD (Not ghost) or B2. I found that the actual shooters elements of the game is dull. Yeah, you got skyhooks and things around you, which are very handy, but the amount of enemies I had to face was annoying and at times, boring. You also can only have 2 weapons at a time. But what I did like was the power ups, which all had different uses. But for half the game I used like 2 of them, which I think they're is like over 7 or something. So why is this game so good? This game really shines, and is utterly unbeaten, with its engaging storyline. What a story, I never thought a game could engage me and make me care about the characters like this. I kept pushing on the average shooter because of this story. Also, there is a lot of freedom for the play, or as it seems. The graphics is more than good and the city is really awesome to, which you get to explore bits and pieces of the downtown dark alley ways to the wealth and good looking streets, parks, houses and more. So overall, this game is a masterpiece in its own right, and has defiantly set a new started in terms of storytelling games. (P.S. The ending will blow you away and I played most of this game on the PS3)8.5/10"

Sumarização baseada em clustering:

"I found that the actual shooters elements of the game is dull.Also, there is a lot of freedom for the play, or as it seems.But for half the game I used like 2 of them, which I think they're is like over 7 or something."

Sumarização baseada em Textrank:

"If you're looking to have fun with mate or just by yourself, play, BF4, COD (Not ghost) or B2.This game really shines, and is utterly unbeaten, with its engaging storyline.Also, there is a lot of freedom for the play, or as it seems."

Ao todo, passamos por 25 análises, das quais expomos duas aqui para análise.

Conclusão

Baseado na review 1, é possível ver a diferença entre os dois algoritmos claramente. O autor da análise enfatiza bastante que com o tempo, atualizações no jogo fizeram com que o jogo demore bastante para iniciar.

O resumo de Clustering aponta outros pontos além da demora para iniciar, essa demora é rapidamente mencionada com a sentença escolhida para representar esse tópico (um grupo conteve vários pontos mas mesmo assim apenas 1 foi escolhido como representante).

Agora o resumo em Textrank passa muito mais tempo enfatizando o fato de que o jogo demora para iniciar, mantendo a ideia de ênfase do autor.

Além disso, o resumo do Clustering deixou um “It’s awesome” no final, dando a entender que o autor recomenda o jogo, mas na verdade o autor não recomenda. O resumo de Textrank já deixa mais claro a opinião negativa do autor.

Na review 2, temos o resumo de clustering com falta de informação, uma das sentenças referência a quantidade de armas disponíveis para o jogador nesse jogo, mas a sentença não faz sentido sem a contraparte, além de falar 100%.

O resumo em Textrank por sua vez está bem mais sucinto, deixa mais claro a opinião do autor, indicando que esse jogo pode nao ser melhor do que outros jogos do mesmo estilo, mas se destaca pela sua história cativante.

Com a análise dos resultados e definições do grupo, concluímos que a sumarização funciona de forma esperada, e em grande parte das vezes gera um resumo condizente e satisfatório, mesmo que às vezes não seja possível inferir a opinião geral do usuário a partir do resumo.

E ainda, os resumos gerados pelo Textrank tendem a refletir melhor o texto original, pelo menos de forma mais precisa e sucinta do que os gerados por Clustering.

De forma geral, percebemos que o algoritmo de sumarização por clustering consegue ressaltar os diferentes "assuntos" principais discutidos dentro do texto,por outro lado, o Textrank foca no principal assunto discutido. A vantagem do resultado do primeiro algoritmo é que ele cobre mais assuntos, a desvantagem é que a construção final do texto resumido parece muito menos coesa do que se comparado ao Textrank, que por se ter assuntos semelhantes, seu resumo aparente ter uma leitura mais fluída. Também tivemos o insight de que o algoritmo de clustering está mais propenso a alterar a opinião inicial do texto, uma vez que ele perde a noção de “peso” dos diferentes assuntos mencionados, e foca em expô-los mesmo que o primeira cluster tenha sido muito mais relevante que os demais.