MO810A : Topicos em Inteligência Artificial

Exercício

23 de novembro de 2016

1 Ponto Principal de um Texto

Em sumarização, costuma-se buscar o ponto principal de uma determinada fonte, ou seja, sua essência (ou *gist*). Identificar essa essência, contudo, é algo bastante subjetivo. Ainda assim, sumarizadores automáticos abordam essa tarefa com base em algumas heurísticas, como as vistas em aula. Estas são, por exemplo:

- Quanto mais frequente uma palavra, mais importante ela é para o assunto tratado no texto. Isso, claro, descontando-se as palavras naturalmente frequentes no português (como "e", "ou", "a", "então" etc) que, via de regra, servem para manter a coerência e coesão textual;
- Em alguns estilos de texto, o título já contém informação relevante, portanto suas palavras possuem uma relevância maior; e
- Em alguns estilos de texto, a informação mais relevante está em seu início, em vez de seu fim.

Um dos estilos de texto que parece obedecer a esse padrão é o texto jornalístico. Dessa forma, uma ferramenta que siga essas heurísticas deveria obter bons resultados nesse estilo de texto. E isso é precisamente o que você irá verificar nesse exercício. Nossa análise não será profunda, mas servirá para uma melhor compreensão das potencialidades da técnica vista em aula quando defronte a exemplos de ordem prática.

2 Metodologia

Para o exercício, você deve coletar 30 textos de notícias (do lugar que preferir), de tamanhos variados. Estes então serão submetidos a um sistema que você irá implementar. O sistema tomará cada texto como entrada, produzindo, em sua saída, a sentença que reflete o ponto principal do texto (ou seja, da notícia).

Para este trabalho, tomaremos como sentença qualquer *string* de texto terminado em '.', '!', '...' e '?'. A cada sentença será atribuído um peso, sendo a de maior peso escolhida então

como ponto principal do texto. Para a definição do peso de uma sentença você deverá agir de forma semelhante ao exemplo dado em aula:

1. Remova do texto todas as palavras mais comuns do português (as *stop words*). Para uma lista das *stop words* a serem consideradas consulte a Seção 5 desse documento. Isso é uma forma de implementar a primeira heurística acima.

2. Faça então a contagem de cada palavra resultante, marcando sua frequência no texto. Palavras existentes no título devem ser incluídas na contagem com peso 2 (ou seja, contam como se fossem 2 palavras). Isso é uma forma de implementar a segunda heurística acima.

3. Para cada sentença (que não seja o título) do texto:

   (a) Associe à sentença um peso calculado da seguinte forma:

   $$P_{\text{sentença}} = \frac{\sum_{i=1}^{n} \text{Frequência}(palavra_i)}{n}$$

   Ou seja, o peso da sentença é a soma das frequências de cada uma de suas palavras no documento (lembrando do peso extra para palavras também contidas no título), dividido pelo número n de palavras da sentença (ambas as contagem feitas após a remoção das *stop words*).

   (b) Multiplique o peso da sentença por $1,1^{1/d}$, onde d é a distância da sentença ao título, contada em número de sentenças (a primeira sentença do texto está a uma distância 1, a segunda a 2 e assim por diante). Isso é uma forma de implementar a terceira heurística acima. Ao final desse passo, a sentença terá um peso correspondente a $P'_{\text{sentença}} = P_{\text{sentença}} \times 1,1^{1/d}$

4. Para o *gist* do texto, retorne a sentença com maior peso ($P'_{\text{sentença}}$). Em caso de empate, escolha a mais próxima do título. Vale notar que o título, embora participe do cálculo de frequência de palavras, não é considerado como uma sentença candidata a *gist*.

3 Avaliação

Para avaliar o sistema, você deve ler as notícias que baixou, definindo manualmente a sentença que reflete o ponto principal do texto. Em seguida, deve rodar seu sistema e comparar a sentença por ele escolhida com a escolhida por você, anotando quantas vezes ele acertou. Para uma melhor discussão de quão bom foi esse resultado, você deve compará-lo ao que se esperaria de acerto caso o sistema escolhesse aleatoriamente uma sentença do texto (para isso, você pode trabalhar com o número médio de sentenças dos textos, produzindo uma taxa média esperada para esse sistema aleatório, e comparando então o resultado com a média de acertos do seu sistema).

Como uma segunda avaliação, e refletindo a subjetividade natural da escolha do *gist* de um texto, indique também se a escolha do sistema pode ser considera aceitável, ou seja, se poderia substituir sua escolha sem maiores problemas.

Verifique também se houve alguma relação da taxa de acerto do *gist*, medida pelos dois métodos acima, com o tamanho do texto analisado.

4 Material para Entrega

Para entrega, você deve fazer um relatório narrando o que foi feito, e apresentando os resultados obtidos em detalhes.

5 Stop Words

Para esse exercício, as seguintes *stop words* devem ser consideradas:

a	à	agora	ainda	alguém	algum
alguma	algumas	alguns	ampla	amplas	amplo
amplos	ante	antes	ao	aos	após
aquela	aquelas	aquele	aqueles	aquilo	as
até	através	cada	coisa	coisas	com
como	contra	contudo	da	daquele	daqueles
das	de	dela	delas	dele	deles
depois	dessa	dessas	desse	desses	desta
destas	deste	deste	destes	deve	devem
devendo	dever	deverá	deverão	deveria	deveriam
devia	deviam	disse	disso	disto	dito
diz	dizem	do	dos	e	é
ela	elas	ele	eles	em	enquanto
entre	era	essa	essas	esse	esses
esta	está	estamos	estão	estas	estava
estavam	estávamos	este	estes	estou	eu
fazendo	fazer	feita	feitas	feito	feitos
foi	for	foram	fosse	fossem	grande
grandes	há	isso	isto	já	la
la	lá	lhe	lhes	lo	mas
me	mesma	mesmas	mesmo	mesmos	meu
meus	minha	minhas	muita	muitas	muito
muitos	na	não	nas	nem	nenhum
nessa	nessas	nesta	nestas	ninguém	no
nos	nós	nossa	nossas	nosso	nossos
num	numa	nunca	o	os	ou
outra	outras	outro	outros	para	pela

pelas	pelo	pelos	pequena	pequenas	pequeno
pequenos	per	perante	pode	pôde	podendo
poder	poderia	poderiam	podia	podiam	pois
por	porém	porque	posso	pouca	poucas
pouco	poucos	primeiro	primeiros	própria	próprias
próprio	próprios	quais	qual	quando	quanto
quantos	que	quem	são	se	seja
sejam	sem	sempre	sendo	será	serão
seu	seus	si	sido	só	sob
sobre	sua	suas	talvez	também	tampouco
te	tem	tendo	tenha	ter	teu
teus	ti	tido	tinha	tinham	toda
todas	todavia	todo	todos	tu	tua
tuas	tudo	última	últimas	último	últimos
um	uma	umas	uns	vendo	ver
vez	vindo	vir	vos	vós	