Checkpoint 1 - *Processamento de Linguagem Natural, chatbots & virtual agents*

Recados paroquiais:

Data máxima de envio: 18/04/2023 23:59

Respostas objetivas, máximo 2 páginas de respostas no total.

Arquivo final em formato .pdf

NOME: Henrico Nardelli Bela

RM: 95985

1 Explique o processo de tokenização em NLP e discuta sua importância na análise de texto. Dê um exemplo de como a tokenização pode favorecer a compreensão semântica de um texto. (Composição da nota do checkpoint: 15%)

R.: A tokenização é um processo de pré-processamento em NLP (Processamento de Linguagem Natural) que consiste em dividir um texto em unidades menores chamadas de "tokens". Esses tokens podem ser palavras, pontuações, números ou até mesmo símbolos especiais. A importância da tokenização na análise de texto é que ela fornece uma base para todas as outras tarefas de processamento de linguagem natural, como análise de sentimento, análise de tópicos, análise de entidades e outras.

A tokenização é importante porque a compreensão semântica do texto depende da capacidade do modelo de linguagem de entender o significado de cada palavra e como elas se relacionam entre si. O processo de tokenização permite que o modelo entenda as palavras individuais, o que é importante para a compreensão do texto. Além disso, a tokenização também é útil para normalizar o texto, tornando-o mais consistente para processamento posterior.

Um exemplo de como a tokenização pode favorecer a compreensão semântica de um texto é o seguinte: imagine que temos a seguinte frase "Eu não gosto de comida italiana." Sem a tokenização, o modelo de linguagem trataria a frase inteira como um único token. No entanto, com a tokenização, o modelo seria capaz de entender que a frase é composta de várias palavras e que "não" é uma palavra negativa que muda o significado da frase inteira. Isso permitiria que o modelo compreendesse melhor o significado da frase e, por sua vez, produzisse resultados mais precisos em tarefas de processamento de linguagem natural, como a análise de sentimentos.

2 Considere o seguinte cenário:

Cenário 1: O texto Alfa possui índice de similaridade de Jaccard (*Jaccard similarity*) de 60% com o texto Beta

A) A partir das informações contidas no cenário 1, é possível inferir que o texto Beta e o texto Alfa tratam de 60 % do mesmo assunto? Sim ou Não e por quê? (Composição da nota do checkpoint: 10%)

R.: A partir das informações contidas no cenário 1, não é possível inferir que o texto Beta e o texto Alfa tratam de 60% do mesmo assunto. A similaridade de Jaccard é uma medida que quantifica a sobreposição entre dois conjuntos, ou seja, a proporção de elementos comuns entre eles. No caso de textos, o conjunto em questão é o conjunto de palavras presentes em cada um dos textos.

Portanto, a similaridade de Jaccard entre dois textos indica apenas a proporção de palavras comuns entre eles, mas não necessariamente o conteúdo ou tema do texto. É possível que dois textos tenham uma alta similaridade de Jaccard, mas abordem tópicos completamente diferentes.

Por exemplo, suponha que o texto Alfa seja sobre "como cozinhar lasanha" e o texto Beta seja sobre "como fazer um bolo de chocolate". Se os dois textos tiverem algumas palavras em comum, como "ovo" e "farinha", a similaridade de Jaccard entre eles poderia ser alta. No entanto, os textos tratam de assuntos completamente diferentes. Portanto, é importante lembrar que a similaridade de Jaccard é apenas uma medida de sobreposição entre conjuntos de palavras e não deve ser interpretada como uma medida de similaridade de conteúdo ou tema entre textos.

B) Explique as diferenças entre a similaridade de cosseno e a similaridade de Jaccard em NLP. Discuta sobre como essas métricas podem ser usadas para agrupar documentos com base em seus conteúdos semelhantes. Dê um exemplo de aplicação dessas métricas para agrupamento de documentos em um conjunto de dados de notícias. (Composição da nota do checkpoint: 20%)

R.: A similaridade de cosseno e a similaridade de Jaccard são duas medidas de similaridade amplamente utilizadas em NLP para comparar a semelhança entre dois documentos.

A similaridade de cosseno mede a semelhança entre dois vetores de termos, onde cada termo corresponde a uma palavra presente nos documentos. Essa medida é dada pelo cosseno do ângulo entre os vetores e varia entre 0 e 1, onde 1 indica uma similaridade perfeita. A similaridade de cosseno é amplamente utilizada para comparar a semelhança entre documentos em uma base de dados, especialmente em casos em que os documentos são grandes e esparsos, como é comum em NLP.

Por outro lado, a similaridade de Jaccard mede a sobreposição entre dois conjuntos de palavras, onde cada conjunto corresponde às palavras presentes nos documentos. Essa medida é dada pela proporção entre o número de palavras em comum entre os conjuntos e o número total de palavras nos conjuntos. A similaridade de Jaccard é especialmente útil quando o foco está na presença ou ausência de palavras, e não no número de ocorrências.

Ambas as métricas podem ser usadas para agrupar documentos com base em seus conteúdos semelhantes. A ideia é que documentos semelhantes tenham uma alta similaridade medida por essas métricas. Uma abordagem comum é usar essas medidas para criar uma matriz de similaridade, em que cada entrada na matriz representa a medida de similaridade entre dois documentos. A partir dessa matriz, é possível aplicar técnicas de agrupamento, como o clustering hierárquico, para agrupar documentos com base em seus conteúdos semelhantes.

Um exemplo de aplicação dessas métricas para agrupamento de documentos em um conjunto de dados de notícias seria o seguinte: suponha que temos um conjunto de dados de notícias que queremos agrupar por tópico. Podemos representar cada notícia como um vetor de termos, onde cada termo corresponde a uma palavra presente na notícia. Em seguida, podemos calcular a similaridade de cosseno ou a similaridade de Jaccard entre cada par de vetores de termos para obter uma matriz de similaridade. Finalmente, podemos aplicar técnicas de agrupamento hierárquico para agrupar as notícias por tópico com base em suas similaridades. Esse processo pode ajudar a organizar grandes conjuntos de dados de notícias e facilitar a busca por informações relevantes.

3 Recall e precisão são duas métricas importantes para avaliar a qualidade de modelos de classificação. Com base nisso, avalie os cenários e responda as perguntas.

Cenário 1: você é o gerente de marketing e a partir de uma amostra de clientes a sua equipe produziu modelos de classificação para encontrar aqueles clientes que possuem probabilidade de consumir o seu produto. Nas métricas de teste, os resultados são os seguintes:

Nome_modelo	PRECISÃO	Clientes Marcados	Erros	Acertos	Recall
1	89.20	426	46	380	13.98
2	87.73	807	99	708	26.05
3	85.31	1089	160	929	34.18
4	83.83	1311	212	1099	40.43
5	82.16	1530	273	1257	46.25
6	80.68	1713	331	1382	50.85
7	78.34	1902	412	1490	54.82
8	76.05	2096	502	1594	58.65
9	73.33	2347	626	1721	63.32

a. Com base no cenário 1, caso você queira maximizar a quantidade de aquisições de seu produto, qual modelo você deve escolher para colocar em produção e abordar clientes? Resposta de 1 a 9 com o porquê. (Composição da nota do checkpoint: 35%)

R.: Com base nas informações dadas, acredito que o melhor modelo a ser utilizado seja o modelo 6, pois ele apresenta um recall e precisão com uma certa harmonia entre ambos, com uma quantidade de acertos e erros condizentes com as métricas, visando obter a melhor probabilidade de que um cliente irá consumir um produto, de acordo com a quantidade de clientes marcados, e usando uma quantidade de amostra mediana entre o mínimo e máximo.

Cenário 2: As métricas abaixo se referem a 9 modelos diferentes que classificam um conjunto de tweets entre “positivo” ou “negativo” em termos de análise de sentimentos. Sendo que essa base de tweets não apresenta outra possibilidade sentimentos além de positivo ou negativo. Seguem as métricas:

Nome_modelo	PRECISÃO	Recall
1	89.20	13.98
2	87.73	26.05
3	85.31	34.18
4	83.83	40.43
5	82.16	46.25
6	80.68	50.85
7	78.34	54.82
8	76.05	58.65
9	73.33	63.32

b. Caso você queira obter pelo menos 50% do total de tweets positivos com a maior precisão possível. Qual modelo escolher? Resposta de 1 a 9 com o porquê. (Composição da nota do checkpoint: 20%)

R.: Modelo 6, visando a maior precisão possível, pois o modelo 6 ainda conta com uma ótima precisão, sem baixar tanto o Recall.