Inteligência Artificial

Thiago Henrique Leite da Silva, RA: 139920

AULA11: Exercício teórico PLN

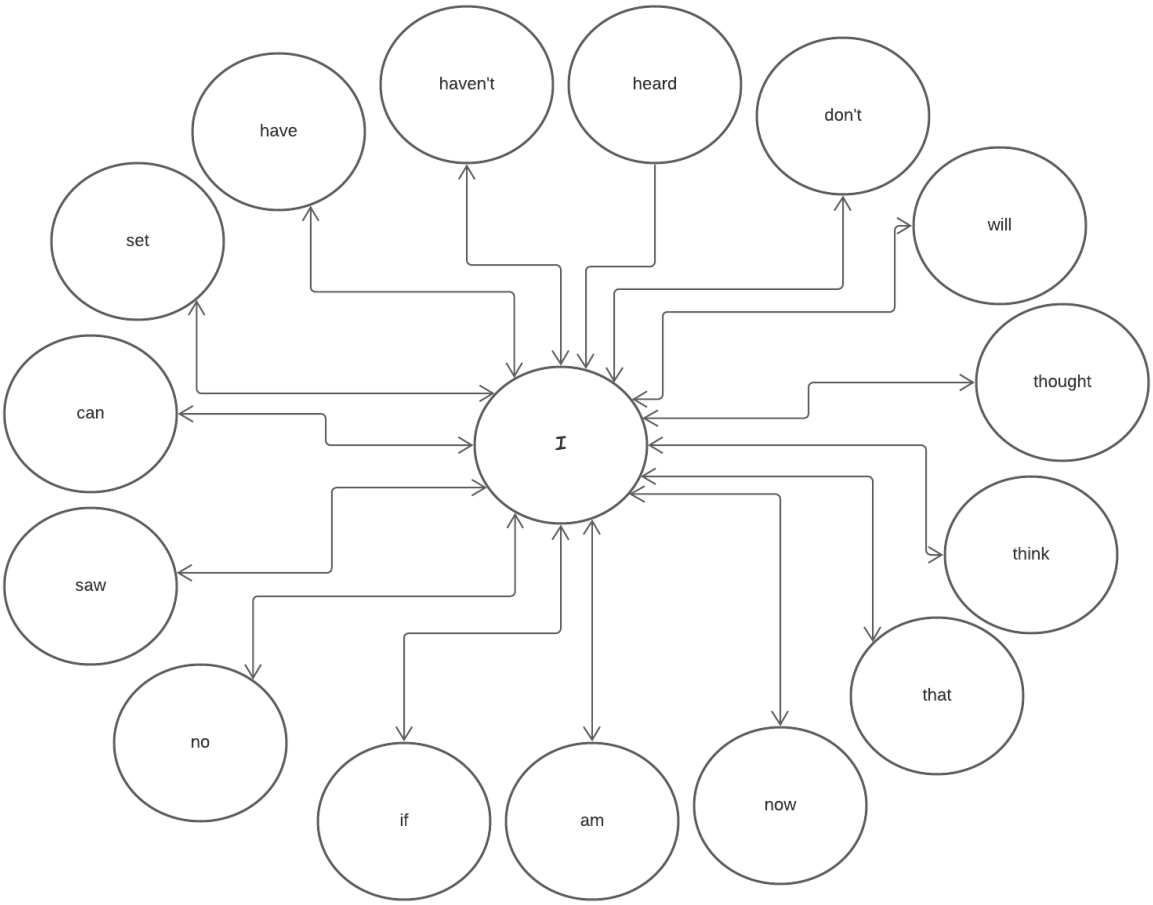
1) Selecione uma música e monte bigramas em forma de grafos com a letra, analise quais palavras mais se repetem. Calcule a probabilidade de cada bigrama ocorrer na letra:

Losing My Religion, R.E.M.

“...

Oh life is bigger
It's bigger than you
And you are not me
The lengths that I will go to
The distance in your eyes
Oh no I've said too much
I set it up
That's me in the corner
That's me in the spot-light
Losing my religion
Trying to keep up with you
And I don't know if I can do it
Oh no I've said too much
I haven't said enough
I thought that I heard you laughing
I thought that I heard you sing
I think I thought I saw you try
Every whisper, of every waking hour
I'm choosing my confessions
Trying to keep an eye on you
Like a hurt, lost and blinded fool, fool
Oh no I've said too much
I set it up
Consider this
Consider this the hint of the century
Consider this the slip
That brought me to my knees, failed
What if all these fantasies come
Flailing around
Now I've said too…

”...



Palavra que mais ocorre = I (Eu)

P(I|have) = 4/30
P(I|haven’t) = 1/30
P(I|will) = 1/30
P(I|set) = 2/30
P(I|don’t) = 1/30
P(I|can) = 1/30
P(I|thought) = 3/30
P(I|heard) = 2/30
P(I|think) = 1/30
P(I|saw) = 1/30
P(I|am) = 1/30
P(I|that) = 3/30
P(I|no) = 3/30
P(I|if) = 1/30
P(I|now) = 1/30

2) Considere o seguinte conjunto de treinamento. Classifique com kNN (k = 1) a sentença:

“I always like foreign films”.

Compare a distância Cosseno com Euclidiana. Pode remover stop words.

Classe	Texto
Negativo (-)	Just plain boring
Negativo (-)	Entirely predictable and lacks energy
Negativo (-)	No surprises and very few laughs
Positivo (+)	Very powerful
Positivo (+)	The most fun film of the summer

Primeiramente montamos o bag of words da frase “I always like foreign films”:

Removeremos a palavra “I” da frase pois é uma stop word.

	just	plain	boring	entirely	predictable	lacks	energy	no	surprise	few	laughs	powerful	most	fun	film	summer	always	like	foreig
T1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T2	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
T3	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0
T4	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
T5	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1

Depois, calculamos as distâncias:

Biblioteca utilizada para o cálculo: https://docs.scipy.org/doc/scipy/reference/spatial.distance.html

- Euclidiana

A distância de S em relação a

- T1: 2.445467343
- T2: 2.646654634
- T3: 2.644676786
- T4: 2.048084800
- T5: 2.623747384

- Cosseno

A distância de S em relação a

- T1: 1.0

- T2: 1.0
- T3: 1.0
- T4: 1.0
- T5: 1.0

Para o vizinho mais próximos portanto, a distância Euclidiana classifica a frase como sendo **positiva**, já em relação a distância Cosseno, ficou indefinido, pois todas as distâncias foram iguais, mostrando assim que, para este caso, ela não se aplica.

3) Considere o seguinte conjunto de treinamento. Classifique com Naive Bayes a sentença: “eu gosto deste lugar”

Classe	Texto
Negativo (-)	eu não gosto deste restaurante
Negativo (-)	estou cansado dessas coisas
Positivo (+)	eu me sinto bem com essas cervejas
Positivo (+)	eu amo esse sanduíche
Positivo (+)	este é um lugar incrível!

Primeiramente vamos classificar as palavras do nosso vocabulário e remover as stop words, feito esses passos, chegamos ao seguinte resultado:

P (positivo)	3/5
P (negativo)	2/5
Vocabulário	Não, Gosto, Restaurante, Cansado, coisas, sinto, bem, cervejas, amo, sanduíche, lugar, incrível
Tamanho do Vocabulário	12 = 7+ e 5-
Stop words	Eu, deste, dessas, me, com, essas, esse, este, é

Agora, para classificar a frase “eu gosto deste lugar”, temos que calcular sua probabilidade de ser positiva e também sua probabilidade de ser negativa, para posteriormente compararmos os resultados:

Removeremos as palavras “eu” e “deste” da frase pois são stop words.

Como a palavra “gosto” não aparece no conjunto de palavras negativas, e a palavra “lugar” não aparece no conjunto de palavras positivas, estamos utilizando o estimador de Laplace, somando uma unidade fictícia para que nossas probabilidades não fiquem zeradas, por isto na hora de efetuarmos a divisão, estamos incrementando duas unidades, que se referem as palavras que adicionamos por conta do estimador, dado que sempre que adicionamos uma palavra em uma classe, adicionamos também na outra.

Sendo assim, calculamos as probabilidades individuais de cada palavra ocorrer em cada uma das classes, que são dadas da seguinte forma:

S = "Eu gosto deste lugar"

P(gosto | +) = (1 + 1) / (9 + 12) P(gosto | -) = (0 + 1) / (7 + 12)

P(lugar | +) = (0 + 1) / (9 + 12) P(lugar | -) = (1 + 1) / (7 + 12)

Por fim, com as probabilidades calculadas, conseguimos efetuar a classificação com o algoritmo Naive Bayes, que fica da seguinte maneira:

$P(+)P(S|+) = 3/5 * ((2*1) / (21^2)) = 0.6 * 0.004535147 = 0.002721088$

$P(-)P(S|-) = 2/5 * ((1*2) / (19^2)) = 0.4 * 0.005540166 = 0.002216066$

Sendo assim, de acordo com nosso algoritmo, a frase “eu gosto deste lugar” pode ser classificada como **positiva**.