Natural Language Processing Principles and Applications (Elective -II) [AML 5233]

Marks: 50 Duration: 90 mins.

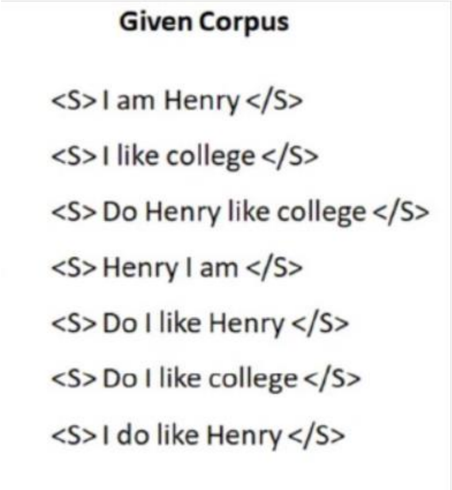
II Sessional Exam – Answer Scheme

What is n-gram? Illustrate the various types and the need of using n-grams in NLP (5 marks)

Estimate the Bi-gram probability? What is the most probable next word predicted by the

model for the following word sequence? (5 Marks)

S I ?



Given Corpus

<S> I am Henry </S>

<S> I like college </S>

<S> Do Henry like college </S>

<S> Henry I am </S>

<S> Do I like Henry </S>

<S> Do I like college </S>

<S> I do like Henry </S>

n-gram: An n-gram is a contiguous sequence of n items from a given sample of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the application. The n-grams typically are collected from a text or speech corpus.

Types of n-gram:

Unigram (1-gram): No history is used.

Bi-gram (2-gram): One word history.

Tri-gram (3-gram): Two words history.

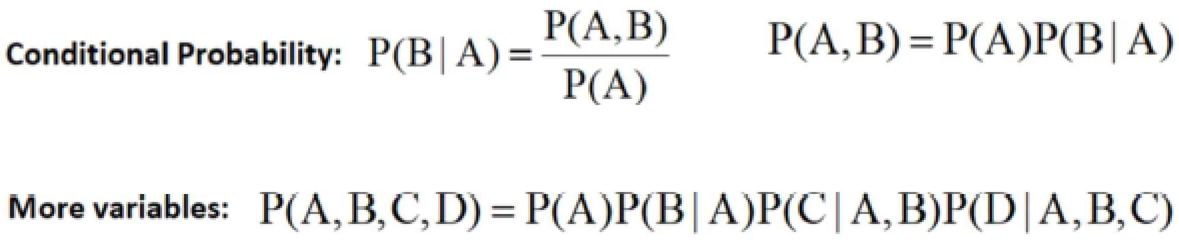
Four-gram (4-gram): Three words history.

Five-gram (5-gram): Four words history is used.

N-grams are a powerful tool for natural language processing (NLP) that can help with a variety of tasks. They can help capture the probability distribution of words in a language, which can be useful for: machine translation, speech recognition, auto-completion, text classification and clustering, and feature engineering.

Word	Frequency
<S>	7
</S>	7
I	6
am	2
henry	5
like	5
college	3
do	4

Next word prediction probability Wi-1 = I.



Conditional Probability: $P(B \mid A) = \dfrac{P(A,B)}{P(A)}$ $\quad P(A,B) = P(A)P(B \mid A)$

More variables: $P(A,B,C,D) = P(A)P(B \mid A)P(C \mid A,B)P(D \mid A,B,C)$

Next word	**Probability of next word =** $\boldsymbol{count(Wi{-}1,\ Wi\ )count(Wi{-}1)count(Wi{-}1,\ Wi\ )count(Wi{-}1)}$
P(</S> | I)	0/6
P(am | I)	2/6
P(Henry | I)	0/6
P(like | I)	3/6
P(college | I)	0/6
P(do | I)	1/6

Probable next word is like as it has the highest value of Probability.

Identify the sentence that gets a higher probability with tri-gram model. (Use the same corpus as in Q1.) (6 Marks)

a) <s> Do I like college </s>

b) <s> Do I like Henry </s>

Comment if there is a need for Laplace smoothing required here. (4 Marks)

Word	***Frequency***
<s>	*7*
</s>	*7*
I	*6*
am	*2*
henry	*5*
like	*5*
college	*3*
do	*4*

a) <s> Do I like college </s>

= P(I|<s> Do) x P(like|Do I) x P(College|I like) x P(</s>|like College)

= 2/3 x 2/2 x 2/3 x 3/3

= 0.44

b) <s> I like Henry </s>

= P(like|<s> I) x P(Henry|I like) x P(</s>|like Henry)

= 1/3 x 1/3 x 2/2

= 0.11

The sentence with higher probability is <s> Do I like Henry </s>

Laplace smoothing prevents the model from assigning zero probabilities to features not present in the training data, ensuring that the model can make predictions for previously unseen words. Since we don't have any zero probabilities here, we don't need to use Laplace Smoothing.

If Laplace Smoothing is to be done, then we add 1 to the numerator of all the probabilities and add the total number of words in the vocabulary to the denominator.

Classify the different relations the words can have with each other in the context of NLP with suitable examples. (10 Marks)

Relations that words can have with each other are:

1. **Synonyms:** Synonyms are words having similar meaning and context and have the same meaning in some or all contexts.

Examples: Filbert/hazelnut**,** Couch/sofa**,** Big/large, Automobile/car, Vomit/throw up, Water/H_2O

Note that there are probably no examples of perfect synonym.

- Even if many aspects of meaning are identical
- Still may differ based on politeness, slang, register, genre, etc.

For example: Water / H20 are synonyms but, it cannot be used interchangeably in sentence like "H2O" in a surfing guide?

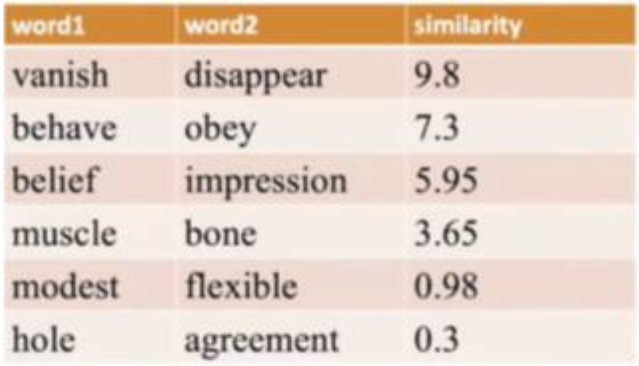
Big / large: are synonyms but, it cannot be used interchangeably in sentence like my big sister as:

my big sister != my large sister

2. **Similarity:** Words with similar meaning. Not synonyms, but sharing some element of meaning.

Example: car-bicycle, Cow-horse

Words similarity can be measured as:



word1	word2	similarity
vanish	disappear	9.8
behave	obey	7.3
belief	impression	5.95
muscle	bone	3.65
modest	flexible	0.98
hole	agreement	0.3

3. **Word Relatedness:** Words can be related in any way, perhaps via a semantic word or field. Any piece of text data—a simple word, sentence, or document—relates back to some natural language.

Example: coffee, tea – are similar words as they both fall under the category of hot beverages.

Coffee, cup – are related words but they cannot be considered similar, as coffee can have an association with the word cup.

4. **Semantic field:** Words that:

- cover a particular semantic domain
- bear structured relations with each other.

The simplest definition of semantics is the study of meaning. Linguistics has its own subfield of linguistic semantics, which deals with the study of meaning in language, the relationships between words, phrases, and symbols, and their indication, meaning, and representation of the knowledge they signify. In simple words, semantics is more concerned with the facial expressions, signs, symbols, body language, and knowledge that are transferred when passing messages from one entity to another.

Example:

Hospitals - surgeon, scalpel, nurse, anaesthetic, hospital are all words that make sense in the context of Hospital.

Restaurants - waiter, menu, plate, food, menu, chef are the words that make sense in the context of a Restaurant.

Houses – likewise door, roof, kitchen, family, bed have some relationship in the context of House.

5. **Antonyms:** Antonyms are pairs of words that define a binary opposite relationship. These words indicate specific sense and meaning that are completely opposite to each other. The state of being an antonym is called antonymy. Senses that are opposite to only one feature of the meaning. Otherwise, they are similar.

Example: Dark/light, Hot/cold, Short/long, fast/slow, up/down, rise/fall, in/out

More formally antonyms can

- Define a binary opposition or be at opposite ends of a scale - long/short, fast/slow
- Be reverses - rise/fall, up/down

6. **Connotation (Sentiment):**

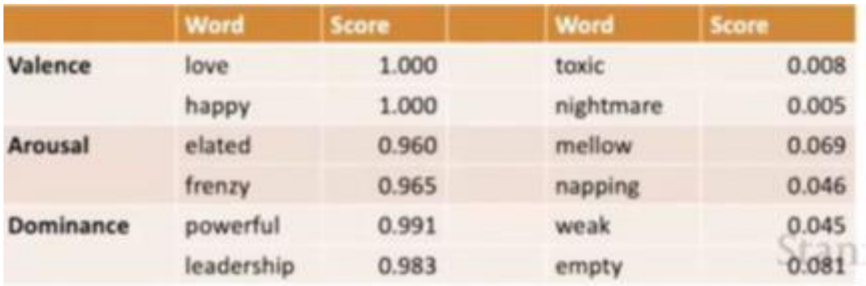
   The word connotation has different meanings in different fields, but here it means the aspects of a word's meaning that are related to a writer or reader's emotions, sentiment, opinions, or evaluations. In addition to their ability to help determine the affective status of a text, connotation lexicons can be useful features for other kinds of affective tasks, and for computational social science analysis.

- Words have affective meanings
- Positive connotations (happy)

- Negative connotations (sad)
- Connotations can be subtle:
- Positive connotation: copy, replica, reproduction
- Negative connotation: fake, knockoff, forgery
- Evaluation (sentiment!)
- Positive evaluation (great, love)
- Negative evaluation (terrible, hate)

Words seem to vary along 3 affective dimensions:

- **Valence:** the pleasantness of the stimulus
- **Arousal:** the intensity of emotion provoked by the stimulus
- **Dominance:** the degree of control exerted by the stimulus



	Word	Score		Word	Score
Valence	love	1.000		toxic	0.008
	happy	1.000		nightmare	0.005
Arousal	elated	0.960		mellow	0.069
	frenzy	0.965		napping	0.046
Dominance	powerful	0.991		weak	0.045
	leadership	0.983		empty	0.081

Define connotation of a word. (2 Marks)

Interpret the different ideas of defining the meaning of a word. (8 Marks)

Connotation Definition

We can define connotation by an associated meaning of a word suggested apart from its explicit or primary meaning. The connotative meaning of a word is based on the shared emotional association with a word. Now, there can be either positive, negative, or neutral connotations. A connotation is an additional meaning to a word.

The connotation is an expression or secondary meaning of a word, which is expressed by a word in addition to its primary meaning. It paints a picture or invokes a feeling. It is created when you mean something else, something that might be initially hidden. Words can be divided into negative, positive, and neutral connotations. A rich vocabulary allows you to choose the right words to express yourself. Choosing the right words is essential while you communicate. Although two words may have the same meaning, their connotations may vary. The words we choose significantly change the meaning of a sentence.

Negative Connotation

The negative connotation also called unfavorable connotation, is the word describing the negative qualities or the disabilities or are disrespectful of a person. It is a bad feeling or negative vibes that people get when hearing a specific word or phrase. It is a word whose connotation implies negative emotions and associations. In a sentence "the aroma of my grandmother's cooking", if we change "aroma" so that it now reads "the stench of my grandmother's cooking," the meaning changes completely. Both "aroma" and "stench" instead of having the same meaning smell, "stench" has a negative connotation, thus, the meal sounds much less appealing.

Logic

By logic, the connotation is roughly synonymous with intention. Connotation often differs from denotation, which is more or less synonymous with extension. Otherwise, the connotation of the word may be thought of as the set of all its possible meanings. The denotation of a word is the collection of things it refers to. Its connotation is what it implies about the things it is used to refer to. The denotation of a dog is (like) a four-legged canine carnivore. Hence saying, "You are a dog" would connote that you were bad rather than denoting you as a canine.

Importance Of Connotation

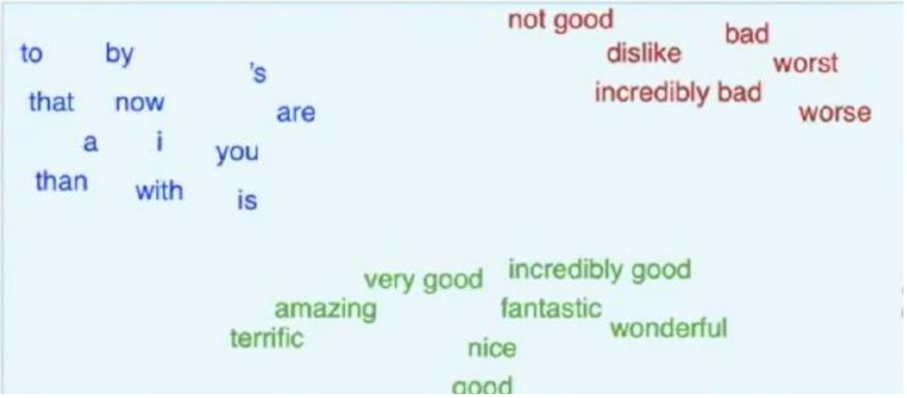
It is important to note that not all are solely 'positive' or 'negative' connotations, depending on how a word is used, it can connote different things. Thus, it is one of the most critical things to consider when it comes to word choice, both in literature and everyday conversation. The emotions or meanings associated with words can be everything. While writing or speaking, connotation places a style to clearly express one's intentions. They can obtain certain emotions or reactions or help to provide distinct impressions of things. Mutually, choosing words with the wrong connotation can produce an undesired reaction or emotion and misrepresent one's intentions.

Defining a meaning by linguistic distribution:

Defining meaning as a point in multidimensional space:

Defining meaning as a point in space based on distribution

Each word is a vector (not just "good" or "w_{45}"). Similar words are "nearby in semantic space". We build this space automatically by seeing which words are nearby in text.



We define meaning of a word as a vector called an “embedding” because it’s embedded into a space.

The standard way to represent meaning in NLP. Every modern NLP algorithm used embeddings as the representation of word meaning. Fine grained model of meaning for similarity.

Consider Sentiment analysis:

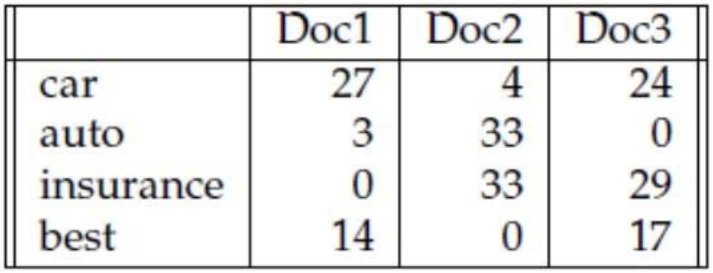
With words, a feature is word identity. Feature 5: the previous word was “terrible” and it requires exact same word to be in training and test set.

With Embeddings:

Feature is a word vector. ‘The previous word was vector [35, 22, 17..]. Now in the test set we might see a similar word vector [34, 21, 14]. We can then generalize it to similar but unseen words.

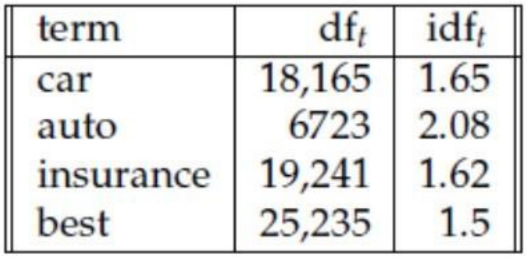
What is tf-idf? Discuss the significance of tf-idf algorithm in NLP. (5 Marks)

Consider the table of term frequencies for 3 documents denoted Doc1, Doc2, Doc3 in the Figure below



	Doc1	Doc2	Doc3
car	27	4	24
auto	3	33	0
insurance	0	33	29
best	14	0	17

Compute the tf-idf weights for the terms car, auto, insurance, best, for each document, using the idf values from Figure below (5 Marks)



term	df_t	idf_t
car	18,165	1.65
auto	6723	2.08
insurance	19,241	1.62
best	25,235	1.5

The co-occurrence matrices above represent each cell by frequencies, either of words with documents or words with other words. But raw frequency is not the best measure of association between words. Raw frequency is very skewed and not very discriminative. If we want to know what kinds of contexts are shared by cherry and strawberry but not by digital and information, we’re not going to get good discrimination from words like the, it, or they, which occur frequently with all sorts of words and aren’t informative about any particular word.

It’s a bit of a paradox. Words that occur nearby frequently (maybe pie nearby cherry) are more important than words that only appear once or twice. Yet words that are too frequent—ubiquitous, like the or good—are unimportant. The tf-idf weighting (the ‘-’ here is a hyphen, not a minus sign) is the product of two terms, each term capturing one of these two intuitions: The first term frequency is the term frequency the frequency of the word t in the document d. We can just use the raw count as the term frequency:

$$\mathrm{tf}_{t,\,d} = \mathrm{count}(\mathrm{t},\,\mathrm{d})$$

More commonly we squashthe raw frequency a bit, by using the log10 of thefrequency instead. The intuition is that a word appearing 100 times in a document doesn't make that word 100 times more likely to berelevant to the meaning of the document. Because we can't take the log of 0, we normally add 1 to the count:

$$\mathrm{tf}_{t,\,d} = \log_{10}(\mathrm{count}(\mathrm{t},\,\mathrm{d})+1)$$

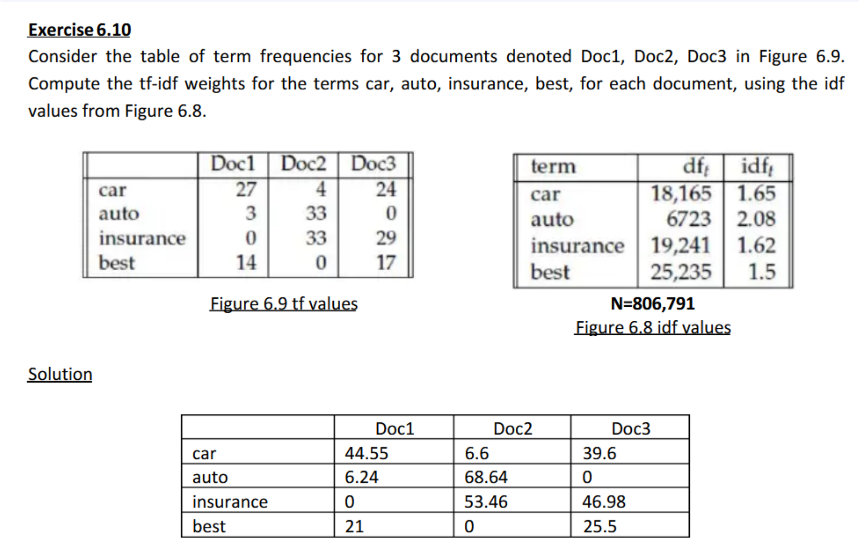
The second factor in *tf-idf* is used to give a higher weight to words that occur only in a few documents. Terms that are limited to few documents are useful for discriminating those documents from the rest of the collection; terms that occur document frequently across the entire collection aren't as helpful. The document frequency df_t of a term t is the number of documents it occurs in. Document frequency is not the same as the collection frequency of a term, which is the total number of times the word appears in the whole collection in any document.

The *idf* is defined using the fraction N/df_t, where N is the total number of documents in the collection, and *dft* is the number of documents in which term t occurs. The fewer documents in which a term occurs, the higher this weight. The lowest weight of 1 is assigned to terms that occur in all the documents. Because of the large number of documents in many collections, this measure too is usually squashed with a log function. The resulting definition for inverse document frequency (idf) is thus

$$idf_t = \log_{10}(N/df_t)$$

The tf-idf weighted value $w_{t,d}$ for word t in document d thus combines term frequency $\mathrm{tf}_{t,d}$ with idf

$$w_{t,d} = tf_{t,d} \times idf_t$$



Exercise 6.10

Consider the table of term frequencies for 3 documents denoted Doc1, Doc2, Doc3 in Figure 6.9. Compute the tf-idf weights for the terms car, auto, insurance, best, for each document, using the idf values from Figure 6.8.

	Doc1	Doc2	Doc3
car	27	4	24
auto	3	33	0
insurance	0	33	29
best	14	0	17

Figure 6.9 tf values

term	df_t	idf_t
car	18,165	1.65
auto	6723	2.08
insurance	19,241	1.62
best	25,235	1.5

N=806,791

Figure 6.8 idf values

Solution

	Doc1	Doc2	Doc3
car	44.55	6.6	39.6
auto	6.24	68.64	0
insurance	0	53.46	46.98
best	21	0	25.5