30 Questions to test a data scientist on Natural Language Processing [Solution: Skilltest – NLP]

SHIVAM BANSAL, JULY 3, 2017

Introduction

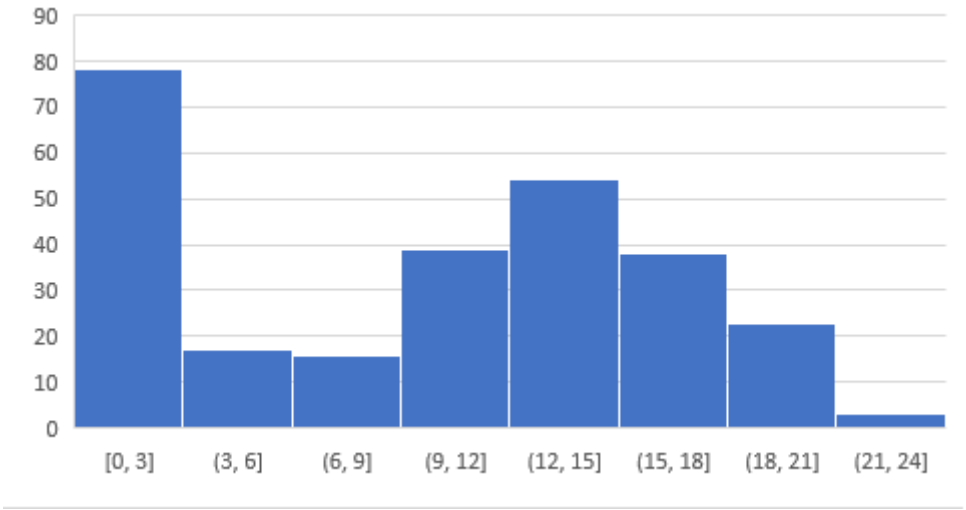
Humans are social animals and language is our primary tool to communicate with the society. But, what if machines could understand our language and then act accordingly? Natural Language Processing (NLP) is the science of teaching machines how to understand the language we humans speak and write.

We recently launched an NLP skill test on which a total of 817 people registered. This skill test was designed to test your knowledge of Natural Language Processing. If you are one of those who missed out on this skill test, here are the questions and solutions.

Here are the leaderboard ranking for all the participants.

Overall Distribution

Below are the distribution scores, they will help you evaluate your performance.



You can access the scores here. More than 250 people participated in the skill test and the highest score obtained was 24.

Helpful Resources

Here are some resources to get in-depth knowledge of the subject.

- Natural Language Processing Made Easy – using SpaCy (in Python)
- Ultimate Guide to Understand & Implement Natural Language Processing (with codes in Python)

Skill Test Questions and Answers

Q1 Which of the following techniques can be used for the purpose of keyword normalization, the process of converting a keyword into its base form?

1. Lemmatization
2. Levenshtein
3. Stemming
4. Soundex

A) 1 and 2
B) 2 and 4
C) 1 and 3
D) 1, 2 and 3
E) 2, 3 and 4
F) 1, 2, 3 and 4

Solution: (C)

Lemmatization and stemming are the techniques of keyword normalization, while Levenshtein and Soundex are techniques of string matching.

2) N-grams are defined as the combination of N keywords together. How many bi-grams can be generated from given sentence:

"Analytics Vidhya is a great source to learn data science"

A) 7
B) 8
C) 9
D) 10
E) 11

Solution: (C)

Bigrams: Analytics Vidhya, Vidhya is, is a, a great, great source, source to, To learn, learn data, data science

3) How many trigrams phrases can be generated from the following sentence, after performing following text cleaning steps:

- Stopword Removal
- Replacing punctuations by a single space

"#Analytics-vidhya is a great source to learn @data_science."

A) 3
B) 4
C) 5
D) 6
E) 7

Solution: (C)

After performing stopword removal and punctuation replacement the text becomes: “Analytics vidhya great source learn data science”

Trigrams – Analytics vidhya great, vidhya great source, great source learn, source learn data, learn data science

4) Which of the following regular expression can be used to identify date(s) present in the text object:

“The next meetup on data science will be held on 2017-09-21, previously it happened on 31/03, 2016”

A) \d{4}-\d{2}-\d{2}
B) (19|20)\d{2}-(0[1-9]|1[0-2])-[0-2][1-9] C) (19|20)\d{2}-(0[1-9]|1[0-2])-([0-2][1-9]|3[0-1])
D) None of the above

Solution: (D)

None if these expressions would be able to identify the dates in this text object.

Question Context 5-6:

You have collected a data of about 10,000 rows of tweet text and no other information. You want to create a tweet classification model that categorizes each of the tweets in three buckets – positive, negative and neutral.

5) Which of the following models can perform tweet classification with regards to context mentioned above?

A) Naive Bayes
B) SVM
C) None of the above

Solution: (C)

Since, you are given only the data of tweets and no other information, which means there is no target variable present. One cannot train a supervised learning model, both svm and naive bayes are supervised learning techniques.

6) You have created a document term matrix of the data, treating every tweet as one document. Which of the following is correct, in regards to document term matrix?

1. Removal of stopwords from the data will affect the dimensionality of data
2. Normalization of words in the data will reduce the dimensionality of data
3. Converting all the words in lowercase will not affect the dimensionality of the data

A) Only 1
B) Only 2
C) Only 3
D) 1 and 2
E) 2 and 3
F) 1, 2 and 3

Solution: (D)

Choices A and B are correct because stopword removal will decrease the number of features in the matrix, normalization of words will also reduce redundant features, and, converting all words to lowercase will also decrease the dimensionality.

7) Which of the following features can be used for accuracy improvement of a classification model?

A) Frequency count of terms
B) Vector Notation of sentence
C) Part of Speech Tag
D) Dependency Grammar
E) All of these

Solution: (E)

All of the techniques can be used for the purpose of engineering features in a model.

8) What percentage of the total statements are correct with regards to Topic Modeling?

1. It is a supervised learning technique
2. LDA (Linear Discriminant Analysis) can be used to perform topic modeling
3. Selection of number of topics in a model does not depend on the size of data
4. Number of topic terms are directly proportional to size of the data

A) 0
B) 25
C) 50
D) 75
E) 100

Solution: (A)

LDA is unsupervised learning model, LDA is latent Dirichlet allocation, not Linear discriminant analysis. Selection of the number of topics is directly proportional to the size of the data, while number of topic terms is not directly proportional to the size of the data. Hence none of the statements are correct.

9) In Latent Dirichlet Allocation model for text classification purposes, what does alpha and beta hyperparameter represent-

A) Alpha: number of topics within documents, beta: number of terms within topics False
B) Alpha: density of terms generated within topics, beta: density of topics generated within terms False
C) Alpha: number of topics within documents, beta: number of terms within topics False
D) Alpha: density of topics generated within documents, beta: density of terms generated within topics True

Solution: (D)

Option D is correct

10) Solve the equation according to the sentence "I am planning to visit New Delhi to attend Analytics Vidhya Delhi Hackathon".

A = (# of words with Noun as the part of speech tag)
B = (# of words with Verb as the part of speech tag)
C = (# of words with frequency count greater than one)

What are the correct values of A, B, and C?

A) 5, 5, 2
B) 5, 5, 0
C) 7, 5, 1
D) 7, 4, 2
E) 6, 4, 3

Solution: (D)

Nouns: I, New, Delhi, Analytics, Vidhya, Delhi, Hackathon (7)

Verbs: am, planning, visit, attend (4)

Words with frequency counts > 1: to, Delhi (2)

Hence option D is correct.

11) In a corpus of N documents, one document is randomly picked. The document contains a total of T terms and the term "data" appears K times.

What is the correct value for the product of TF (term frequency) and IDF (inverse-document-frequency), if the term "data" appears in approximately one-third of the total documents?

A) KT * Log(3)
B) K * Log(3) / T
C) T * Log(3) / K
D) Log(3) / KT

Solution: (B)

formula for TF is K/T

formula for IDF is log(total docs / no of docs containing "data")

= log(1 / (⅓))

= log (3)

Hence correct choice is Klog(3)/T

Question Context 12 to 14:

Refer the following document term matrix

Term	Document						
	d1	*d2*	*d3*	*d4*	*d5*	*d6*	*d7*
t1	2	1	0	0	0	0	0
t2	1	2	0	0	0	0	1
t3	3	1	0	0	1	1	0
t4	0	0	1	2	1	1	1
t5	0	0	1	1	1	1	1
t6	0	0	1	1	0	0	0

12) Which of the following documents contains the same number of terms and the number of terms in the one of the document is not equal to least number of terms in any document in the entire corpus.

A) d1 and d4
B) d6 and d7
C) d2 and d4
D) d5 and d6

Solution: (C)

Both of the documents d2 and d4 contains 4 terms and does not contain the least number of terms which is 3.

13) Which are the most common and the rarest term of the corpus?

A) t4, t6
B) t3, t5
C) t5, t1
D) t5, t6

Solution: (A)

T5 is most common terms across 5 out of 7 documents, T6 is rare term only appears in d3 and d4

14) What is the term frequency of a term which is used a maximum number of times in that document?

A) t6 – 2/5
B) t3 – 3/6
C) t4 – 2/6
D) t1 – 2/6

Solution: (B)

t3 is used max times in entire corpus = 3, tf for t3 is 3/6

15) Which of the following technique is not a part of flexible text matching?

A) Soundex
B) Metaphone
C) Edit Distance
D) Keyword Hashing

Solution: (D)

Except Keyword Hashing all other are the techniques used in flexible string matching

16) True or False: Word2Vec model is a machine learning model used to create vector notations of text objects. Word2vec contains multiple deep neural networks

A) TRUE
B) FALSE

Solution: (B)

Word2vec also contains preprocessing model which is not a deep neural network

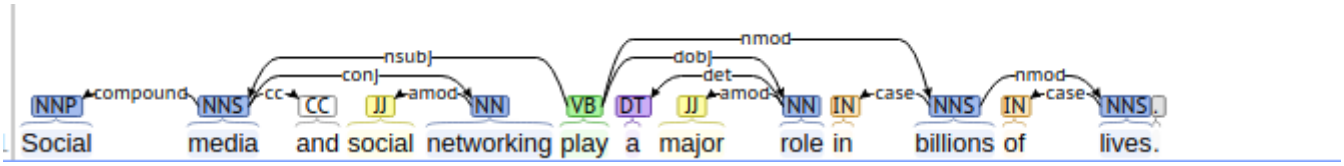
17) Which of the following statement is(are) true for Word2Vec model?

A) The architecture of word2vec consists of only two layers – continuous bag of words and skip-gram model
B) Continuous bag of word is a shallow neural network model
C) Skip-gram is a deep neural network model
D) Both CBOW and Skip-gram are deep neural network models
E) All of the above

Solution: (C)

Word2vec contains the Continuous bag of words and skip-gram models, which are deep neural nets.

18) With respect to this context-free dependency graphs, how many sub-trees exists in the sentence?



A) 3
B) 4
C) 5
D) 6

Solution: (D)

Subtrees in the dependency graph can be viewed as nodes having an outward link, for example:

Media, networking, play, role, billions, and lives are the roots of subtrees

19) What is the right order for a text classification model components

1. Text cleaning
2. Text annotation
3. Gradient descent
4. Model tuning
5. Text to predictors

A) 12345
B) 13425
C) 12534
D) 13452

Solution: (C)

A right text classification model contains – cleaning of text to remove noise, annotation to create more features, converting text-based features into predictors, learning a model using gradient descent and finally

tuning a model.

20) Polysemy is defined as the coexistence of multiple meanings for a word or phrase in a text object. Which of the following models is likely the best choice to correct this problem?

A) Random Forest Classifier
B) Convolutional Neural Networks
C) Gradient Boosting
D) All of these

Solution: (B)

CNNs are popular choice for text classification problems because they take into consideration left and right contexts of the words as features which can solve the problem of polysemy

21) Which of the following models can be used for the purpose of document similarity?

A) Training a word 2 vector model on the corpus that learns context present in the document
B) Training a bag of words model that learns occurrence of words in the document
C) Creating a document-term matrix and using cosine similarity for each document
D) All of the above

Solution: (D)

word2vec model can be used for measuring document similarity based on context. Bag Of Words and document term matrix can be used for measuring similarity based on terms.

22) What are the possible features of a text corpus

1. Count of word in a document
2. Boolean feature – presence of word in a document
3. Vector notation of word
4. Part of Speech Tag
5. Basic Dependency Grammar
6. Entire document as a feature

A) 1
B) 12
C) 123
D) 1234
E) 12345
F) 123456

Solution: (E)

Except for entire document as the feature, rest all can be used as features of text classification learning model.

23) While creating a machine learning model on text data, you created a document term matrix of the input data of 100K documents. Which of the following remedies can be used to reduce the dimensions of data –

1. Latent Dirichlet Allocation
2. Latent Semantic Indexing
3. Keyword Normalization

A) only 1
B) 2, 3
C) 1, 3
D) 1, 2, 3

Solution: (D)

All of the techniques can be used to reduce the dimensions of the data.

24) Google Search's feature – "Did you mean", is a mixture of different techniques. Which of the following techniques are likely to be ingredients?

1. Collaborative Filtering model to detect similar user behaviors (queries)
2. Model that checks for Levenshtein distance among the dictionary terms
3. Translation of sentences into multiple languages

A) 1
B) 2
C) 1, 2
D) 1, 2, 3

Solution: (C)

Collaborative filtering can be used to check what are the patterns used by people, Levenshtein is used to measure the distance among dictionary terms.

25) While working with text data obtained from news sentences, which are structured in nature, which of the grammar-based text parsing techniques can be used for noun phrase detection, verb phrase detection, subject detection and object detection.

A) Part of speech tagging
B) Dependency Parsing and Constituency Parsing
C) Skip Gram and N-Gram extraction
D) Continuous Bag of Words

Solution: (B)

Dependency and constituent parsing extract these relations from the text

26) Social Media platforms are the most intuitive form of text data. You are given a corpus of complete social media data of tweets. How can you create a model that suggests the hashtags?

A) Perform Topic Models to obtain most significant words of the corpus
B) Train a Bag of Ngrams model to capture top n-grams – words and their combinations
C) Train a word2vector model to learn repeating contexts in the sentences
D) All of these

Solution: (D)

All of the techniques can be used to extract most significant terms of a corpus.

27) While working with context extraction from a text data, you encountered two different sentences: The tank is full of soldiers. The tank is full of nitrogen. Which of the following measures can be used to remove the problem of word sense disambiguation in the sentences?

A) Compare the dictionary definition of an ambiguous word with the terms contained in its neighborhood
B) Co-reference resolution in which one resolute the meaning of ambiguous word with the proper noun present in the previous sentence
C) Use dependency parsing of sentence to understand the meanings

Solution: (A)

Option 1 is called Lesk algorithm, used for word sense disambiguation, rest others cannot be used.

28) Collaborative Filtering and Content Based Models are the two popular recommendation engines, what role does NLP play in building such algorithms.

A) Feature Extraction from text
B) Measuring Feature Similarity
C) Engineering Features for vector space learning model
D) All of these

Solution: (D)

NLP can be used anywhere where text data is involved – feature extraction, measuring feature similarity, create vector features of the text.

29) Retrieval based models and Generative models are the two popular techniques used for building chatbots. Which of the following is an example of retrieval model and generative model respectively.

A) Dictionary based learning and Word 2 vector model
B) Rule-based learning and Sequence to Sequence model
C) Word 2 vector and Sentence to Vector model
D) Recurrent neural network and convolutional neural network

Solution: (B)

choice 2 best explains examples of retrieval based models and generative models

30) What is the major difference between CRF (Conditional Random Field) and HMM (Hidden Markov Model)?

A) CRF is Generative whereas HMM is Discriminative model
B) CRF is Discriminative whereas HMM is Generative model
C) Both CRF and HMM are Generative model
D) Both CRF and HMM are Discriminative model

Solution: (B)

Option B is correct

End Notes

I tried my best to make the solutions as comprehensive as possible but if you have any questions/doubts please drop in your comments below. And I would love to hear the feedback about the skill test. Feel free to share them in comments below. For latest and upcoming skill test please refer to the DataHack platform of Analytics Vidhya.

If you want to learn more about Natural Language Processing and how it is implemented in Python, then check out our video course on NLP using Python.

Happy Learning!