Text Processing

Feed the following paragraph into your favourite data analytics tool, and answer the following:

As a term, data analytics predominantly refers to an assortment of applications, from basic business intelligence (BI), reporting and online analytical processing (OLAP) to various forms of advanced analytics. In that sense, it's similar in nature to business analytics, another umbrella term for approaches to analyzing data -- with the difference that the latter is oriented to business uses, while data analytics has a broader focus. The expansive view of the term isn't universal, though: In some cases, people use data analytics specifically to mean advanced analytics, treating BI as a separate category. Data analytics initiatives can help businesses increase revenues, improve operational efficiency, optimize marketing campaigns and customer service efforts, respond more quickly to emerging market trends and gain a competitive edge over rivals -- all with the ultimate goal of boosting business performance. Depending on the particular application, the data that's analyzed can consist of either historical records or new information that has been processed for real-time analytics uses. In addition, it can come from a mix of internal systems and external data sources. At a high level, data analytics methodologies include exploratory data analysis (EDA), which aims to find patterns and relationships in data, and confirmatory data analysis (CDA), which applies statistical techniques to determine whether hypotheses about a data set are true or false. EDA is often compared to detective work, while CDA is akin to the work of a judge or jury during a court trial -- a distinction first drawn by statistician John W. Tukey in his 1977 book Exploratory Data Analysis. Data analytics can also be separated into quantitative data analysis and qualitative data analysis. The former involves analysis of numerical data with quantifiable variables that can be compared or measured statistically. The qualitative approach is more interpretive -- it focuses on understanding the content of non-numerical data like text, images, audio and video, including common phrases, themes and points of view.

a. What is the probability of the word “data” occurring in each line?

First, we need to calculate the number of lines. A ready-available Python NLP toolkit called NLTK library is used because simple scripting with regex by considering ending punctuation marks cannot accurately determine the end of a sentence (e.g. John W. Tukey fulfils the grammatical rule of ending punctuation (.) followed by a capital first letter is not the end of a sentence). Then, we calculate the number of words in each line and the occurrence of the word "data" in each line. From there we can get the probability by dividing the occurrence of the word "data" with the number of words in each line.

The probability of the word “data” occurring in each line:

line 1: 0.034

line 2: 0.056

line 3: 0.037

Available at https://github.com/aplatyps/data_science_demo

line 4: 0

line 5: 0.038

line 6: 0.067

line 7: 0.119

line 8: 0

line 9: 0.143

line 10: 0.058

line 11: 0.034

b. What is the distribution of distinct word counts across all the lines?

To calculate the distribution of distinct word count, the same NLP toolkit can be usedto tokenize and count the distribution. However, before that, the text has to be standardised so that the NLP toolkit will not treat singular/plural or capital/small letter words as unique words. This is a process called lemmatization. In this case, I'm using a Python library called pattern. The result is slightly better than WordNet but still have some mistakes. Further post-processing can be done to correct the mistakes for deployment purposes, but for the demo I shall stop here.

The distribution of distinct word counts across all the lines:

'data': 18

'a': 12

'to': 11

'the': 11

'analytic': 10

'of': 10

'and': 9

'be': 9

'in': 6

'analysi': 6

'can': 5

'busines': 4

'that': 4

'or': 4

'term': 3

Available at https://github.com/aplatyps/data_science_demo

'with': 3

'application': 2

'from': 2

'bi': 2

'process': 2

'advance': 2

'for': 2

'approach': 2

'analyze': 2

'used': 2

'while': 2

'have': 2

'focus': 2

'view': 2

'separate': 2

'market': 2

'more': 2

'on': 2

'it': 2

'include': 2

'exploratory': 2

'eda': 2

'which': 2

'cda': 2

'compare': 2

'work': 2

'qualitative': 2

'predominantly': 1

'refer': 1

'an': 1

'assortment': 1

Available at https://github.com/aplatyps/data_science_demo

'basic': 1

'intelligence': 1

'report': 1

'online': 1

'analytical': 1

'olap': 1

'variou': 1

'form': 1

'sense': 1

"it'": 1

'similar': 1

'nature': 1

'another': 1

'umbrella': 1

'difference': 1

'latter': 1

'orient': 1

'broader': 1

'expansive': 1

'universal': 1

'though': 1

'some': 1

'case': 1

'people': 1

'use': 1

'specifically': 1

'mean': 1

'treat': 1

'category': 1

'initiative': 1

'help': 1

Available at https://github.com/aplatyps/data_science_demo

'business': 1

'increase': 1

'revenue': 1

'improve': 1

'operational': 1

'efficiency': 1

'optimize': 1

'campaign': 1

'customer': 1

'service': 1

'effort': 1

'respond': 1

'quickly': 1

'emerge': 1

'trend': 1

'gain': 1

'competitive': 1

'edge': 1

'over': 1

'rival': 1

'all': 1

'ultimate': 1

'goal': 1

'boost': 1

'performance': 1

'depend': 1

'particular': 1

"that'": 1

'consist': 1

'either': 1

'historical': 1

Available at https://github.com/aplatyps/data_science_demo

'record': 1

'new': 1

'information': 1

'real-time': 1

'addition': 1

'come': 1

'mix': 1

'internal': 1

'system': 1

'external': 1

'source': 1

'at': 1

'high': 1

'level': 1

'methodology': 1

'aim': 1

'find': 1

'pattern': 1

'relationship': 1

'confirmatory': 1

'apply': 1

'statistical': 1

'technique': 1

'determine': 1

'whether': 1

'hypothese': 1

'about': 1

'set': 1

'true': 1

'false': 1

'often': 1

Available at https://github.com/aplatyps/data_science_demo

'detective': 1

'akin': 1

'judge': 1

'jury': 1

'dure': 1

'court': 1

'trial': 1

'distinction': 1

'first': 1

'draw': 1

'by': 1

'statistician': 1

'john': 1

'w': 1

'tukey': 1

'hi': 1

'1977': 1

'book': 1

'also': 1

'into': 1

'quantitative': 1

'former': 1

'involve': 1

'numerical': 1

'quantifiable': 1

'variable': 1

'measure': 1

'statistically': 1

'interpretive': 1

'understand': 1

'content': 1

Available at https://github.com/aplatyps/data_science_demo

'non-numerical': 1

'like': 1

'text': 1

'image': 1

'audio': 1

'video': 1

'common': 1

'phrase': 1

'theme': 1

'point': 1

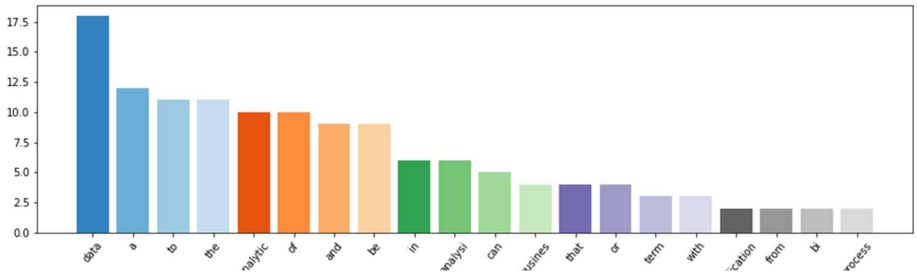

Figure. 1. Distribution of the top 20 most occurring words plotted with frequency binning.

c. What is the probability of the word “analytics” occurring after the word “data”?

Using a similar methodology to (a), the probability is calculated by dividing the occurrence of the word “analytics” after the word “data” and the occurrence of the word "data" in the entire paragraph.

The probability of the word “analytics” occurring after the word “data”:

0.467

=== END OF DOCUMENT ===

Available at https://github.com/aplatyps/data_science_demo