tweets3

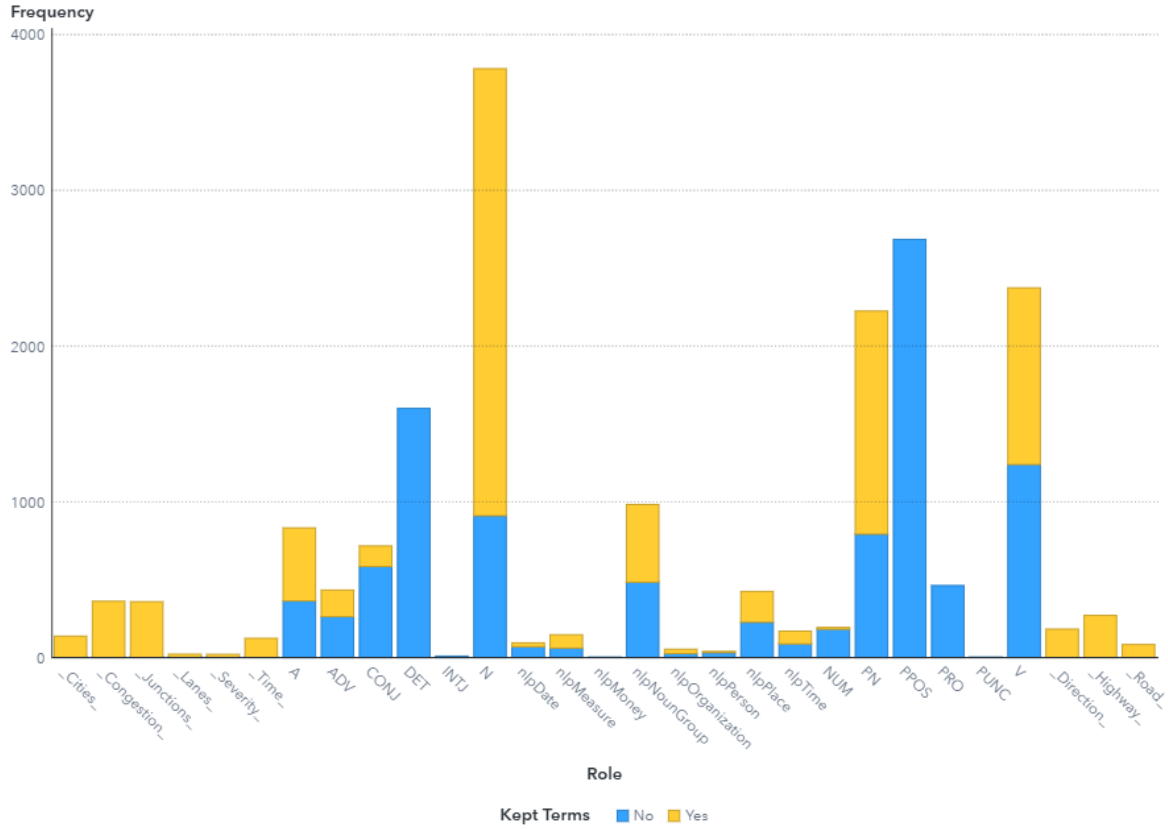
"Text Parsing" Results

by: ta01468@surrey.ac.uk

Contents

Role by Frequency 3
Descriptive Statistics 5

Role by Frequency



The Role by Frequency report is a visual summary of the terms data. Parsing assigns role labels to each term in the entire data set. The height of the bars in this report indicates the frequency for each type of role. In this data set, the most frequently occurring roles are N (noun) and PPOS (preposition/postposition). Together they represent 34.25% of the terms. Note that both bare numbers and most types of punctuation (except some symbols) are removed by default from the terms list and are not reported here. To work with the table of values further, download the data.

For each bar, the areas defined by the two colors represent the proportion of kept or dropped terms with the given role. A downstream Topics node will use only the kept terms to determine topics across the data set. In this data set, 53.82% of the terms that are found in the data have been dropped from consideration in building topics. The dropped terms were removed because of their presence in a stop list or because they do not meet the frequency threshold set by the ‘minimum number of documents’ parameter. Additional terms can be dropped in the interactive view or by adding them to the applied stop list.

Parsing identifies terms through a sequence of NLP (natural language processing) steps, including tokenization, multiword identification, lemmatization, part-of-speech tagging, and noun group extraction. Synonym lists and misspelling detection can be

added to the analysis to further group term variants under a single parent term. Concept extraction may also be applied using a preceding Concepts node.

Each resulting role is either a part-of-speech (content or function word) or a concept. Parts-of-speech include labels such as N (noun), CONJ (conjunction), PN (proper noun), ADV (adverb), NUM (numeric), PUNC (punctuation), and so on. Concepts may include noun groups (nlpNounGroup) and those defined and passed forward from a preceding Concepts node in the pipeline.

Descriptive Statistics

Measure	Terms in a Sentence	Terms in a Document
Minimum	1	1
Maximum	52	66
Mean	14.5624	31.6087

The Descriptive Statistics table records patterns that are found across the data set. The average length of sentences by count of terms in this data set is 14.56. The range of sentence length is 1 - 52 terms. The average length of documents by a count of terms in this data set is 31.61. The range of document length is 1 - 66 terms.

The information in this chart can identify unexpected data characteristics that may need to be investigated. The information presented in the table is a summary view only. For more detailed information or to compare data sets, run the profileText action.