ITMO UNIVERSITY

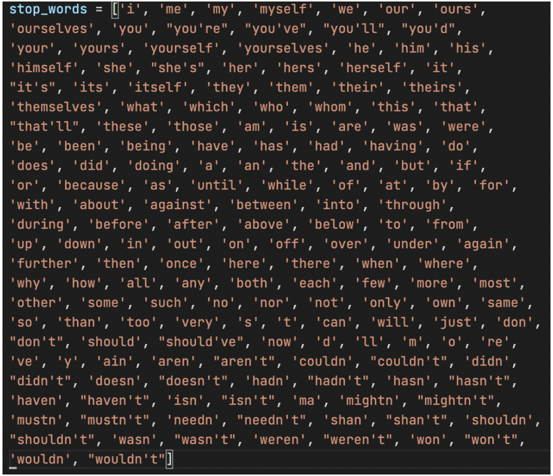
Lab 1 & 2

Sentiment Analysis of Microblog Data Streams

Simon Naumov, M4138

Using prepared helpful data

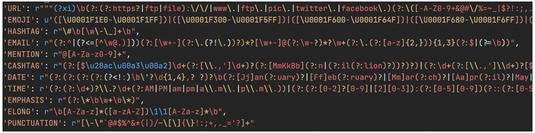
- *Natural Language Toolkit* library with words, punctuation and stop words



```
stop_words = ['i', 'me', 'my', 'myself', 'we', 'our', 'ours',
'ourselves', 'you', "you're", "you've", "you'll", "you'd",
'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his',
'himself', 'she', "she's", 'her', 'hers', 'herself', 'it',
"it's", 'its', 'itself', 'they', 'them', 'their', 'theirs',
'themselves', 'what', 'which', 'who', 'whom', 'this', 'that',
"that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were',
'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do',
'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if',
'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for',
'with', 'about', 'against', 'between', 'into', 'through',
'during', 'before', 'after', 'above', 'below', 'to', 'from',
'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again',
'further', 'then', 'once', 'here', 'there', 'when', 'where',
'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most',
'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same',
'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don',
"don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're',
've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn',
"didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't",
'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't",
'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn',
"shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't",
'wouldn', "wouldn't"]
```

Using prepared helpful data

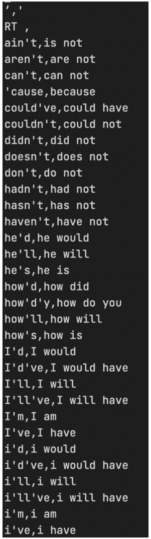
- *Natural Language Toolkit* library with words, punctuation and stop words
- *Regular expressions* for urls, emojis, hashtags, emails and many more metadata



```
'URL': r"""(?xi)\b(?:(?:https?|ftp|file):\/\/|www\.|ftp\.|pic\.|twitter\.|facebook\.)(?:\([-A-Z0-9+&@#\/%=~_|$?!:;,.
'EMOJI': u'([\U0001F1E0-\U0001F1FF])|([\U0001F300-\U0001F5FF])|([\U0001F600-\U0001F64F])|([\U0001F680-\U0001F6FF])|(
'HASHTAG': r"\#\b[\w\-\_]+\b",
'EMAIL': r"(?:^|(?<=[^\w@.)]))(?:[\w+-](?:\.(?!\.))?)*?[\w+-]@(?:\w-?)*?\w+(?:\.(?:[a-z]{2,})){1,3}(?:$|(?=\b))",
'MENTION': r"@[A-Za-z0-9]+",
'CASHTAG': r"(?:[$\u20ac\u00a3\u00a2]\d+(?:[\\.,']\d+)?(?:[MmKkBb](?:n|(?:il(?:lion)?))?)?)|(?:\d+(?:[\\.,']\\d+)?[$
'DATE': r"(?:(?:(?:(?:(?<!:)\b\'?\d{1,4},? ?)?\b(?:[Jj]an(?:uary)?|[Ff]eb(?:ruary)?|[Mm]ar(?:ch)?|[Aa]pr(?:il)?|May|
'TIME': r'(?:(?:\d+)?\\.?\d+(?:AM|PM|am|pm|a\\.m\\.|p\\.m\\.))|(?:(?:[0-2]?[0-9]|[2][0-3]):(?:[0-5][0-9])(?::(?:[0-5
'EMPHASIS': r"(?:\*\b\w+\b\*)",
'ELONG': r"\b[A-Za-z]*([a-zA-Z])\1\1[A-Za-z]*\b",
'PUNCTUATION': r"[\-\"`@#$%^&*(|)/~\[\]{\}!:;+,._='?]+"
```

Using prepared helpful data

- *Natural Language Toolkit* library with words, punctuation and stop words
- *Regular expressions* for urls, emojis, hashtags, emails and many more metadata
- *Contraction* and *emoticons* mappings



```
’,'
RT ,
ain't,is not
aren't,are not
can't,can not
'cause,because
could've,could have
couldn't,could not
didn't,did not
doesn't,does not
don't,do not
hadn't,had not
hasn't,has not
haven't,have not
he'd,he would
he'll,he will
he's,he is
how'd,how did
how'd'y,how do you
how'll,how will
how's,how is
I'd,I would
I'd've,I would have
I'll,I will
I'll've,I will have
I'm,I am
I've,I have
i'd,i would
i'd've,i would have
i'll,i will
i'll've,i will have
i'm,i am
i've,i have
```

Using prepared helpful data

- *Natural Language Toolkit* library with words, punctuation and stop words
- *Regular expressions* for urls, emojis, hashtags, emails and many more metadata
- *Contraction* and *emoticons* mappings

normalized string

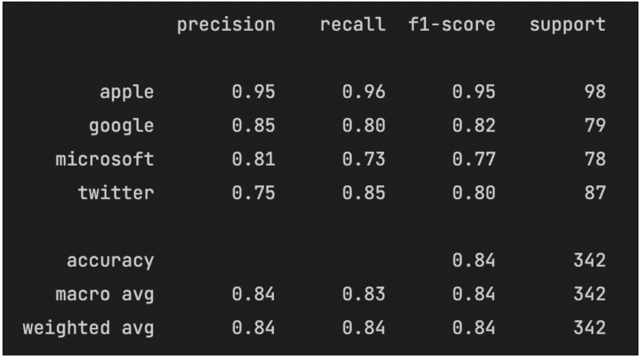
Application

Machine learning algorithm application with the transformed data input

In particular, train and test *Linear Support Vector Classification*

Results

- Organization Prediction



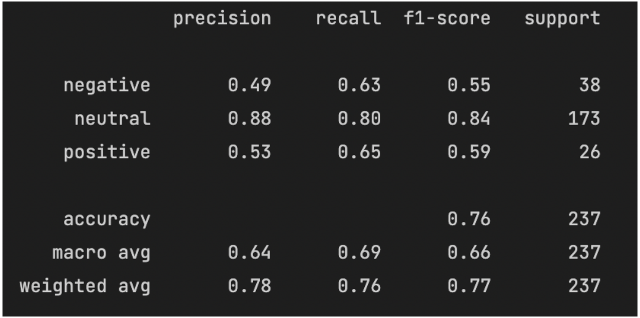
```
              precision    recall  f1-score   support

       apple       0.95      0.96      0.95        98
      google       0.85      0.80      0.82        79
   microsoft       0.81      0.73      0.77        78
     twitter       0.75      0.85      0.80        87

    accuracy                           0.84       342
   macro avg       0.84      0.83      0.84       342
weighted avg       0.84      0.84      0.84       342
```

Results

- Organization Prediction
- Sentiment Analysis



```
              precision    recall  f1-score   support

    negative       0.49      0.63      0.55        38
     neutral       0.88      0.80      0.84       173
    positive       0.53      0.65      0.59        26

    accuracy                           0.76       237
   macro avg       0.64      0.69      0.66       237
weighted avg       0.78      0.76      0.77       237
```

Thank you for your attention!

IT'sMOre than a UNIVERSITY