Stray Concepts

P. S. Langeslag

n-Gram

A sequence consisting of *n* words as they occur in a string of text.

n-Gram

A sequence consisting of n words as they occur in a string of text.

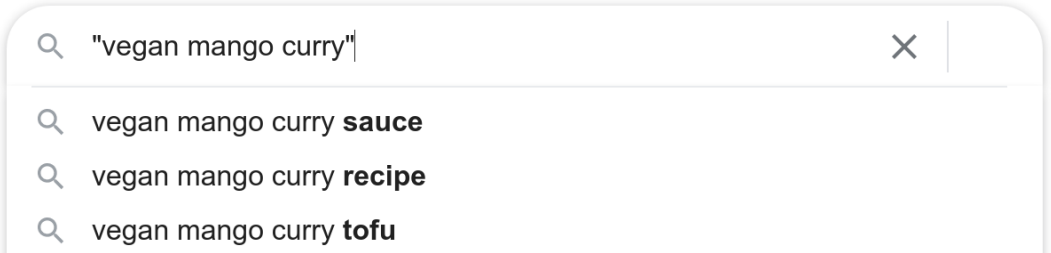

Figure 1: Double quotes yield n-grams on most search engines

- We speak of bigrams and trigrams but commonly write 2-gram, 3-gram.

Regular Expression

Search string relying on an extensive, conventional pattern-matching grammar

```
>>> import re
>>> haystack = "thesis2022-04-19q_formatted.md"
>>> needle = "^thesis[0-9]{4}-[0-9]{2}-[0-9]{2}[a-z]_formatted\..*$"
>>> re.search(needle, haystack)
<re.Match object; span=(0, 30), match='thesis2022-04-19q_formatted.md'>
```

Stem

Linguistic Definition

The base of a given word form, to which inflectional information is added.

Stem

Linguistic Definition

The base of a given word form, to which inflectional information is added.

NLP Definition

The base to which a given type may be reduced by stripping away (known) inflectional (and sometimes derivational) information, whether or not the resulting form is linguistically recognized.

```
>>> import re
>>> sentence = 'Jael rushed hurtling down the stairs'
>>> tokens = sentence.split()
>>> pattern = '(s|ing|ed)$'
>>> stems = [re.sub(pattern,'',token) for token in tokens]
>>> stems
['Jael', 'rush', 'hurtl', 'down', 'the', 'stair']
```

Lemma

Linguistic Definition

Dictionary headword

Lemma

Linguistic Definition

Dictionary headword

NLP Definition

Unique identifier to which inflected forms of the same word may be assigned

Overfitting

Training a supervised neural network so precisely on its training data that its ability to predict new data is adversely affected.

Data Separation

- fitting

- selecting optimal hyperparameters

- demonstrating accuracy with new data

Shuffle your data!

Precision and Recall 1/2

Recall

How well a classifier does at assigning the accurate label: $\frac{TP}{TP + FN}$

Precision and Recall 1/2

Recall

How well a classifier does at assigning the accurate label: $\frac{TP}{TP + FN}$

Precision

How well a classifier does at foregoing assignment of an inaccurate label: $\frac{TP}{TP + FP}$

Precision and Recall 1/2

Recall

How well a classifier does at assigning the accurate label: $\frac{TP}{TP + FN}$

Precision

How well a classifier does at foregoing assignment of an inaccurate label: $\frac{TP}{TP + FP}$

F-Measure

The harmonic mean between the two: $\frac{2rp}{r + p}$

By default, all three are used of a specific label, but they can be generalized.

Precision and Recall 2/2: Example

If a classifier tries to compile a list of e.g. Latin words as found in an English corpus, we discern two kinds of error: **false negatives** are Latin words in the corpus that are not added to the list, whereas **false positives** are non-Latin word that are added to the list.

Recall measures false negatives; it describes how good the algorithm is at finding all the Latin words.

Precision measures false positives; it describes how good the algorithm is at avoiding populating the list with non-Latin words.

Evaluation should capture both these measurements.

Bibliography

Bird, Steven, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly, 2009. https://www.nltk.org/book/.

Eisenstein, Jacob. *Introduction to Natural Language Processing*. Cambridge, MA: MIT Press, 2019.

Jurafsky, Dan, and James H. Martin. *Speech and Language Processing*. 3rd ed. draft., 2021. http://web.stanford.edu/~jurafsky/slp3/.

Lane, Hobson, Cole Howard, and Hannes Hapke. *Natural Language Processing in Action: Understanding, Analyzing, and Generating Text with Python*. Shelter Island, NY: Manning, 2019.

Matthes, Eric. *Python Crash Course*. 2nd ed. San Francisco, CA: No Starch, 2019.

Python Software Foundation. "Python," October 4, 2020. https://www.python.org/.

Vasiliev, Yuli. *Natural Language Processing Using Python and spaCy: A Practical Introduction*. San Francisco: No Starch Press, 2020.