Natural Language Processing

Introduction

Glossary

Chapter X
Edit Distancing, Phonetic Algorithms and N-Grams

Basic Vector Rules
Vector values are either 0 or positive, multiplying a vector increases its magnitude rather than direction, and the triangle inequality holds: the shortest distance between any two points is found in a direct, straight line.

Basic Vector Rules and the Minowski Distance:
Minowski Distance:
A generalized algorithm for finding the distance between data points points. Also known as Lp norm distance calculator. Often, using lower p values, like p = 1 or p = 2, has better results for high dimensional problems.

$(\sum_{i=1}^{n} |x_i - y_i|^p)^{1/p}$ p -> 1: gives you Manhattan Distance p -> 2: gives you Euclidean Distance p -> ∞: gives you Chebychev Distance

Manhattan Distance
It measures the distance between two points by travelling along orthogonal points . It is often better at dealing with high-dimensionality data than Euclidean distance is.

*I*magine travelling in Manhattan city: since there are buildings everywhere, you couldn't travel in a straight line from building A to building B.

$|(x_1 - x_2)| + |(y_1 - y_2)|$

Euclidean Distance:
Fact Checker! A simple way of finding how similar two objects are, by using their x and y attributes, and finding Pythagorean distance between the two points

$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$

Cosine Similarity:
This finds the similarity between two data points.

Cosine Similarity:
EDIT DISTANCING
This is a way of finding the minimum number of edits to be made on a word before it matches with another particular word.

USAGES
It can be used for spell-checking, OCR (optical character recognition) and for finding approximate matches between words, as well as finding the similarity between DNA sequences in Bioinformatics.

1 **LEVENSHTEIN**
This is a way of finding how dissimilar two words are, providing the minimum number of edits to be made on a word before it matches with another particular word.

2 LONGEST COMMON SUBSEQUENCE (LCS)

This is a way of finding the minimum number of edits to be made on a word before it matches with another particular word.

3 HAMMING DISTANCE

This is used to find the distance between categorical values.

4 DAMERAU–LEVENSHTEIN DISTANCE

This is a way of finding the minimum number of edits to be made on a word before it matches with another particular word.

5 JARO DISTANCE

This is a way of finding the minimum number of edits to be made on a word before it matches with another particular word.

6 BAG DISTANCE

This is a cost-effective way of getting an approximate edit distance.

PHONETIC ALGORITHMS

This is a way of encoding text to a phonetic representation. Give general description, high-level summary, so that content can point to ideas beyond just the next list.

SOUNDEX

METAPHONE & DOUBLE METAPHONE

PHONEX & PHONIX

Chapter Y
Named Entity Recognition

NAMED ENTITY RECOGNITION
This provides a way of identifying specific entities, such as people, organisations, and objects, within unstructured data, such as natural text in a news article.

Chapter Z
Syntactical Tools

GRAMMAR INDUCTION
This is using machine learning to generate a formal set of rules about how language is be structured i.e description of a syntax.

LEMMATIZATION
This is a way of breaking down a word into its true base form i.e strpping the prefix and suffix of the word.

STEMMING
This is a way of breaking down a word into a base form i.e stripping the prefix and suffix of the word. However, unlike Lemmatization, this "base form" may not be a real word.

MORPHOLOGICAL (TEXT) SEGMENTATION
This is the act of separating words into individual units of grammar, morphemes. Since English has a fairly simple morphology, i.e no complicated word inflections, it is possible to model all possible forms of a word as separate words.

PARTS OF SPEECH TAGGING (POS)
This is the act of determining which part of speech a word is, e.g subject, object, noun, adjective, adverb, conjunction, etc.

PARSING
This is breaking down a string or sentence into a Tree of Nodes, in other words, using grammatical analysis to identify syntax.

1 **DEPENDENCY PARSING**
This parsing method is simpler than Constituency Parsing

2 **CONSTITUENCY PARSING**
3 **CONTEXT-FREE GRAMMAR (PCFG)**
4 **STOCHASTIC GRAMMAR**

Chapter K
Representation and Vectorization

TOKENIZATION
STOP-WORDS
It is not always beneficial to filter out stop words, because some times they give important context.

BAG OF WORDS
TERM-FREQUENCY INVERSE-DOCUMENT FREQUENCY
This gives a measure of how importance words are to a particular document.

ZIPF'S LAW
The frequency of a word is inversely proportional to its rank in a given document*

Chapter U
Transformers Models**

ATTENTION MECHANISM
This ignores order of words in a sentence and instead converts each word in the sentence into both a column and a record, showing any correlation between each word, ignoring the order of words.

TRANSFORMER MODELS
Can be bi-directional n-grams..

Chapter V
Applied methods

UPSTREAM-DOWNSTREAM TASKS
This is when a model is trained on a specific set of vectors, e.g an upstream task, and is then used on tasks that consist of a different set of input vectors, i.e it's a downstream task.

GENETIC ALGORITHMS
Used to reduce dimensionality when finding matches in large datassets

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \quad \textbf{Equation } 1.$$

This is the appropriate way to display an equation.

```
\begin{equation}
x= \dfrac{-b \pm \sqrt{b^2 - 4ac}}{2a} \label{equ1}
\end{equation}
```

Using the $$ method, the `\[, \]` environment, and the `equation*` environment produce unnumbered equations and should be avoided.

In multiline equations, label only the last line.

$$\begin{aligned} x &= \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \\ &= 2 \pm 3i \end{aligned}$$ **Equation** 2.

This template automatically loads usual `amsthm` and `amsmath` packages. Additional packages should be loaded in the preamble.

CONCLUSIONS

Describe major outcomes, novelty, and significance of your work. Future work may be noted.

ACKNOWLEDGEMENTS

This section is optional. The authors thank and not "would like to thank" such and such an organization or person. Co-authors should not be listed here.

REFERENCES

1. Marquez, V., Frohlich, T., Armache, J. P., Sohmen, D., Donhofer, A., Mikolajka, A., Berninghausen, O., Thomm, M., Beckmann, R., Arnold, G. J., and Wilson, D. N. (2011) Proteomic characterization of archaeal ribosomes reveals the presence of novel archaeal-specific ribosomal proteins, *J Mol Biol* 405, 1215–1232. *https://doi.org/10.33697/ajur.2019.003*
2. Fierke, C. A., and Hammes, G. G. (1996) Transient Kinetic Approaches to Enzyme Mechanisms, in *Contemporary Enzyme Kinetics and Mechanism* (Purich, D., Ed.) 2nd ed., 1–35, Academic Press, New York.
3. Agricultural Research Service, U.S.D.A. National Nutrient Database for Standard Reference, Release 26, *http://ndb.nal.usda.gov/ndb/search/list* (accessed Mar 2014)

ABOUT THE STUDENT AUTHOR

John Smith and Jane Smith will graduate in . . . , *etc.*

NLP REFERENCES

https://web.stanford.edu/ jurafsky/slp3/ https://bib.dbvis.de/uploadedFiles/155.pdf (Distance Metrics)