Natural Language Processing Techniques for Creating Large Language Models

Wendon Doswell and Clay Jones

CS Undergraduate Students and Research Assistants for Cybersecurity Complex

Supported by ACTION AI, NSF Grant Number: 2229876. Advised by Dr. Fields

Abstract

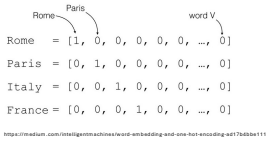
Machines do not understand raw text, text needs to be represented numerically through vectors. Natural Language Processing (NLP) is a branch of artificial intelligence (AI) that trains machine learning (ML) models to understand, generate, and manipulate human language. This research explored NLP techniques to create word vectors and word embeddings to train ML models.

Background

Data modeling techniques include:

One-Hot Encoding

Converts data into a binary vector representation where each word has its own vector



https://medium.com/intelligentmachines/word-embedding-and-one-hot-encoding-ad17bddbe111

Bag-of-Words (BoW)

Counts number of times a word appears in a document to find relevance

	it	is	puppy	cat	pen	a	this
it is a puppy	1	1	1	0	0	1	0
it is a kitten	1	1	0	0	0	1	0
it is a cat	1	1	0	1	0	1	0
that is a dog and this is a pen	0	2	0	0	1	2	1
it is a matrix	1	1	0	0	0	1	0

https://medium.com/rptsmaogh/sterm-frequency-inverse-document-frequency-tf-tbsuzxd26da

TF-IDF

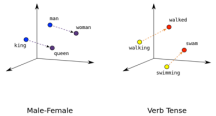
Finds importance of a word by assigning a frequency-based weight

Words	TF (for A)	TF (for B)	IDF	TFIDF (A)	TFIDF (B)
Jupiter	1/5	0	ln(2/1) = 0.69	0.138	0
Is	1/5	1/8	ln(2/2) = 0	0	0
The	1/5	2/8	ln(2/2) = 0	0	0
largest	1/5	0	ln(2/1) = 0.69	0.138	0
Planet	1/5	1/8	ln(2/2) = 0	0.138	0
Mars	0	1/8	ln(2/1) = 0.69	0	0.086
Fourth	0	1/8	ln(2/1) = 0.69	0	0.086
From	0	1/8	ln(2/1) = 0.69	0	0.086
Sun	0	1/8	ln(2/1) = 0.69	0	0.086

https://www.analyticsvidhya.com/blog/2021/07/word2vec-for-word-embeddings-a-beginners-guide/

Word2Vec

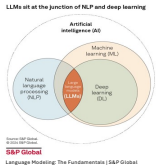
Generate word embedding to find the sematic relationship between words



Word-embeddings in 2020. Review with code examples | by Rostyslav Neskorozhenyi | Towards Data Science

Methodologies

1. Explore the connection between NLP, AI, and LLMs by researching machine learning, LLM creation, and NLP algorithms.
2. Explore background basics of NLP
   Data Preparation – Prepare Corpus
   Data Cleansing – Tokenization, Stop Words
3. Explore Data Modeling Techniques
   One-hot encoding - Bag of Words – TF-IDF – Word2vec



Language Modeling: The Fundamentals | S&P Global

Text Representation

Corpus:

- Doc #0: Norfolk State University Loves Natural Language Processing.
- Doc #1: Norfolk State Spartan Loves Artificial Intelligence.
- Doc #2: Large Language Modes are built Using Natural Language Processing.
- Doc #3: Spartans for NLP.

Bag of Words

	are	artificial	built	for	intelligence	language	large	loves	models	natural	nlp	norfolk	processing	spartans	state	university	using
Doc0	0	0	0	0	0	1	0	1	0	1	0	1	1	0	1	1	0
Doc1	0	1	0	0	1	0	0	1	0	0	0	1	0	0	1	1	0
Doc2	1	0	1	0	0	2	1	0	1	1	0	0	1	0	0	0	1
Doc3	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0

Bag of Words Vector Representation for "Language": [1, 0, 2,1]

TF-IDF

Vocab	are	artificial	built	for	intelligent	language	large	loves	models	natural	nlp	norfolk	processing	spartan	spartans	state	university	using
Doc0	0	0	0	0	0	.3626	0	.3626	0	.3626	0	.3626	.3626	0	0	.3626	.4598	0
Doc1	0	.4534	0	0	.4534	0	0	.3575	0	0	0	.3575	0	.4534	0	.3575	0	0
Doc2	.3385	0	.3385	0	0	.5337	.3385	0	.3385	.2668	0	0	.2668	0	0	0	0	.3385
Doc3	0	0	0	.5774	0	0	0	0	0	0	.5774	0	0	0	.5774	0	0	0

TF-IDF Vector Representation for "Language": [.3626, 0, .5337,0]

Word2Vec

Skip Gram model trained using Google New's 3-billion-word corpus dataset
Context words generated from a Target Word

```
model.most_similar('College')
```

```
[('Collge', 0.6822724938392639),
 ('University', 0.6669971346855164),
 ('Col_lege', 0.6303771138191223),
 ('Univeristy', 0.6178331971168518),
 ('Community_College', 0.6173429489135742),
 ('Unviersity', 0.5917240977287292),
 ('Univer_sity', 0.5827249884605408),
 ('Univerity', 0.5738999843597412),
 ('Colege', 0.5729310512542725),
 ('Colllege', 0.5718283653259277)]
```

```
model.most_similar('Norfolk')
```

```
[('Suffolk', 0.6745657920837402),
 ('Dorset', 0.6166641116142273),
 ('Essex', 0.6098291277885437),
 ('Yarmouth', 0.6097761988639832),
 ('Great_Yarmouth', 0.6018956899642944),
 ('Lowestoft', 0.6018952131271362),
 ('Del._Algie_Howell', 0.5961890816688538),
 ('Cornwall', 0.5862817168235779),
 ('Chichester', 0.585571289062 5),
 ('Lincolnshire', 0.5795595645904541)]
```

Results/Findings

- 60 to 80 percent of training AI models is spent with data preparation and cleansing
- One-Hot – Simplest form, doesn't capture relationship
- BoW - Doesn't capture full relationship between words
- TF-IDF - Evaluate the importance of a word but not the semantics
- Word2vec – gives semantics but not the contextual relationship between words
- Transformers – provides context and semantic relationships between words

Future Work

Explore data modeling with transformer architecture such as

- Bidirectional Encoding Representation from Transformers (BERT) - Google
- Generative Pretrained Transformers (GPT) OpenAI - ChatGPT
- Build a LLM for Cybersecurity

References

[1] S. Arnold, "How to become an Accredited Investor on Linqto," *Private Equity Investing | Linqto Private Investing*, Mar. 27, 2024. What is Artificial Intelligence (AI) and Why it Matters (linqto.com)(accessed Jun. 05, 2024).

[2]"Bag of Words Model in NLP Explained," *Built In*, 2023. https://builtin.com/machine-learning/bag-of-words (accessed Jun. 05, 2024).

[3] IBM, "What is machine learning?," IBM.com. https://www.ibm.com/topics/machine-learning

[7] "Python | Word Embedding using Word2Vec," *GeeksforGe awcbvsgeks*, May 18, 2018. https://www.geeksforgeeks.org/python-word-embedding-using-word2vec/

[8] A. Verma, "Understanding CBOW vs. Skip-gram in Word Embeddings," *Medium*, Nov. 06, 2023. https://ai.plainenglish.io/understanding-cbow-vs-skip-gram-in-word-embeddings-2d2f679dd755?gi=9a4995edfeea (accessed Jun. 20, 2024).

Tools Used

colab GENSIM topic modelling for humans scikit-learn Machine Learning with Scikit-Learn