Week 1 Exploring NLP Applications

Nate

Real-World NLP Applications

Machine Translation

- Machine Translation (MT) is the automated process of translating text or speech from one natural language to another. Its purpose is to bridge communication gaps across different cultures and regions without requiring a human translator for every interaction.
- Google Translate is an example of Machine Translation. It utilizes Neural Machine Translation (NMT) to provide instant translations for many different languages, supporting web pages, documents, and real-time voice conversations.
- NLP enables this by identifying grammatical structures, idiomatic expressions, and context in the source language to produce fluent output in the target language.

Chatbots and Conversational Agents

- These applications are designed to simulate human-like conversation through text or voice. Their purpose is to provide 24/7 customer support, assist with information retrieval, or offer personal companionship.
- Customer support chatbots like Intercom or Zendesk bots are used by e-commerce websites to handle routine inquiries like "Where is my order?" or "How do I reset my password?
- Through Natural Language Understanding (NLU), the system identifies the user's intent and extracts key entities like an order number.

Vectorization with Bag of Words

Vocabulary Extraction

The unique vocabulary within the sentences: a, brother, cause, god, is, likes, Loki, mischief, of, powerful, the, Thor, to

BoW Count Vector Table

Sentence	a	brother	cause	god	is	likes	Loki	mischief	of	powerful	the	Thor	to
S1	1	0	0	1	1	0	0	0	0	1	0	1	0
S2	0	1	0	0	1	0	1	0	1	0	1	1	0
S3	0	0	1	0	0	1	1	1	0	0	0	0	1

Limitations of Bag of Words

- BoW treats sentences as an unordered collection. "Thor defeated Loki" and "Loki defeated Thor" would look identical to the model.
- Common "stop words" like "is" or "the" carry the same weight as high-value words like "mischief."
- As the vocabulary grows, the vectors become very long and mostly contain zeros, which is inefficient.
- It cannot recognize that "powerful" and "strong" have similar meanings.

TF-IDF Improvement

Term Frequency–Inverse Document Frequency (TF-IDF) improves upon BoW by weighing words based on their uniqueness across the entire dataset rather than just their frequency in a single sentence. Using the sentences from Task 2:

- If "Loki" appears twice in a paragraph, it is considered important for that paragraph.
- If the word "is" appears in every single sentence in a book, its importance is "penalized" because it doesn't help us distinguish one sentence from another.

In our example, TF-IDF would mathematically reduce the weight of the word "is" (since it appears in multiple sentences) and increase the weight of the word "mischief" (since it is unique to Sentence 3). This makes the numerical representation much more descriptive of the actual content.