Ref:https://kavita-ganesan.com/fasttext-vs-word2vec/
https://taksw222.tistory.com/188

What is FastText?

another effective method for **representing words numerically (Word Embedding)**, alongside Word2Vec. It was specifically developed by Facebook to overcome some of the limitations of Word2Vec.

Background

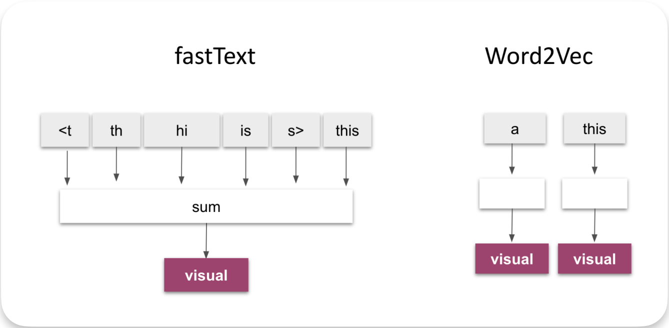
Word2Vec learns word representations based on a given large text corpus. While highly effective, it faces a significant issue: if a new word (Out-Of-Vocabulary, OOV) appears that was not present in the training data, it cannot generate an embedding (numerical representation) for that word. This means that unknown words can cause "errors" or be treated as "unknown," limiting their usability.

Core Idea

The most significant difference between FastText and Word2Vec lies in **how they process words.**

- **Word2Vec:** Learns words by treating the entire word as a single unit. (e.g., learns "apple" as a whole).
- **FastText: Breaks down words into multiple smaller subword units (n-grams) and learns from these.**

For example, if the word is "apple," FastText might decompose it into 3-grams like ap, ppl, ple (if n=3), or 4-grams like appl, pple, etc., and learn from these subwords



Advantages:

1. **Addresses OOV (Out-Of-Vocabulary) Problem:** Even if a word like "uncommon" was not explicitly trained, FastText likely learned embeddings for its constituent subwords such as un, comm, on. It can then **infer** the embedding for "uncommon" by combining the embeddings of these subword units.
2. **Improved Handling of Rare Words:** Even for words that appear infrequently, their constituent subwords might appear in other words. This allows FastText to learn more accurate embeddings for rare words.
3. **Incorporates Morphological Information:** By learning from internal word structures (prefixes, suffixes, roots, etc.), FastText can better capture the relationships between morphologically related words that have different forms but similar meanings, such as "running," "ran," and "runs."