Word2Vec and Doc2Vec

Introduction

One of the biggest tasks in computing is knowledge representation: how do we represent data in a way that is optimal for a computer to utilize it? This fundamental question is generally overlooked, but can make or break an application. Consider the simple task of texting your mother: how would you send that information to them? This isn't something we think about, we would text them in a our native language. Why? There are other human languages, there's Morse code, and a million other ways to express that information, but we choose our native language because we know that it is the representation of the information that is most useful in our scenario.

We could send our mothers a text in some language we know they don't understand, she could translate it, and understand it, but it unlikely she would get the full meaning of what we were trying to say to her. Something would be "lost in translation." Why does this happen? Natural languages send information as sequences of words, but we are trying to convey emotions, thoughts, and other incredibly complex concepts through them. We aren't trying to express to our mothers a series of words, we are trying to express a complex idea using the semantic meanings that the words represent.

This is what makes Natural Language Processing so incredibly difficult: how do we teach a computer the semantic meaning of a sequence of characters? A semantic representation means that we can no longer think about words as atomic units because we need to be able to compare the similarity of the semantic meanings of words [1].