Pranith Bottu
pbottu2
CS410: Text Information Systems
November 7, 2021

Technology Review: Gensim

Pop culture has always been more than just a means of entertainment. Oftentimes, it inspires us to innovate and invent in ways we never quite thought possible. A clear example of this is Star Trek. In 1987, during *Star Trek: The Next Generation*, the Enterprise crew often used gadgets called "Personal Access Data Devices" (PADDs) to input coordinates for star systems. Roughly 23 years later, Apple launched their very first iPad, which was very similar in design to the PADDs. Another instance of pop culture inspiring us can be seen in the concept of an Artificial Intelligence (AI) capable of complex information retrieval in seconds. This was seen in *iRobot* and *Iron Man* when their respective AIs scoured vast datasets to retrieve largely text-based information for their user. Today, we call that Natural Language Processing (NLP) and have an entire industry built around it. While we haven't quite reached that same level of sophistication and efficiency displayed in movies today, we've already started taking steps to reach there. One such step is Gensim.

Gensim is an open-source Python library designed for "topic modeling". It allows us to extract semantic topics from text documents in an efficient manner very painlessly. This is done by first requiring words (tokens) to be converted to unique IDs using a dictionary object. This dictionary is later used to create a "bag of words" corpus that's inputted into Gensim's specialized models along with it. While this process may seem intimidating, built-in Python functions such as *Dictionary.doc2bow()* make it very straightforward. Moving past this, Gensim also offers the ability to create different models following different representations. For example, Term Frequency - Inverse Document Frequency (TFIDF) matrices can be created by leveraging *models.TfidfModel()* from Gensim. This function lets us determine TFIDF weights from our corpus through a simple call. In this same vein of thought, Gensim's *model* package offers a wide variety of similar tools for things like Latent Dirichlet Allocation (LDA) and Latent Semantic Indexing (LSI). At this point, it's evident that Gensim has a lot to offer that we could use today.

One clear use case is summarizing what a text document is about. This is especially important because it saves us the time taken to read a document in its entirety. As of October 1 of this year, the Internet is estimated to contain 5 million terabytes of data. When using search engines, Gensim has a clear application here since it allows us to quickly determine a document's topic and its relevance to a search query. Another example is seen in advertisements and other forms of media meant to attract audiences. News apps such as the New York Times and other popular papers often have large sources of material for viewers to consume. As such, when viewers use their websites, the search bars often use topic modeling to identify the relevance of different articles. The common trend here is that search engines benefit greatly from Gensim. In today's world, there are companies that take advantage of this.

Some companies that utilize Gensim include insurance tech companies like Convr that deal with underwriting. The underwriting process often deals with needing to sift through numerous reports to identify a client's risk and how to quote their insurance plea. To speed up this process, topic modeling tools like Gensim allow us to quickly summarize those said documents and reduce the underwriting time to a fraction of what it was before. Other companies that use Gensim include Avito (advertising) and MailMine (AI in mailbox). For advertising companies like Avito, understanding customer reviews is often very valuable when determining how to better attract customers. That said, those reviews can be numerous. By utilizing Gensim, we can treat those reviews as text documents and quickly summarize their contents. With regards to MailMine's purpose, our email inboxes are often filled to the brim with messages from different areas of our lives. By utilizing topic modeling, we can summarize them quickly and sift through them quickly to gather the main points as opposed to wasting time going through hundreds of emails. Another industry where Gensim is useful is in Customer Support/Services. For training purposes and general information mining, conversations with customers are often recorded and transferred to text documents for future reference. In this use case, topic handling can be used in a different manner to judge the overall tone of those conversations along with what features of a product got the most complaints. By using Gensim, this whole process is sped up. The common pattern here is that these businesses rely upon Gensim for quickly categorizing different documents based on their topics.

Put simply, Gensim provides a very straightforward but vital function: simplifying topic modeling. It does this by providing built-in functions within its packages that simplify complicated processes to simple function calls. This is especially useful for applications such as search engines, customer support, advertising, underwriting, and any other industry reliant on consuming large volumes of text to make decisions. The wide assortment of models in its library makes Gensim relevant and something to consider using in any NLP application. As such, Gensim truly is a step in the right direction towards reaching the same level of text retrieval demonstrated in futuristic pop culture.

References:

- "10 Times Star Trek Predicted the Future (and Other Amazing Sci Fi Inventions That Became Scientific Fact)." *InterFocus Lab Furniture*, 30 Apr. 2020, https://www.mynewlab.com/blog/10-times-star-trek-predicted-future-sci-fi/.
- Brendan McGuigan Last Modified Date: October 01. "How Big Is the Internet?" *EasyTechJunkie*, https://www.easytechjunkie.com/how-big-is-the-internet.htm#:~:text=Eric%20Schmidt%2C%20the%20CEO%20of,data%2C%20or%205%20trillion%20megabytes.
- "Gensim." *Wikipedia*, Wikimedia Foundation, 11 May 2021, https://en.wikipedia.org/wiki/Gensim.
- Li, Zhi. "A Beginner's Guide to Word Embedding with Gensim word2vec Model." *Medium*, Towards Data Science, 1 June 2019, https://towardsdatascience.com/a-beginners-guide-to-word-embedding-with-gensim-word2vec-model-5970fa56cc92.
- Prabhakaran, Selva. "Gensim Tutorial - A Complete Beginners Guide." *Machine Learning Plus*, 13 Oct. 2021, https://www.machinelearningplus.com/nlp/gensim-tutorial/.
- "Why Developers like Gensim." *StackShare*, https://stackshare.io/gensim.