Zero to NLP Hero: Functional Introduction to Natural Language Processing and Semantic Embeddings

Description

Excited by Large Language Models like ChatGPT but have no idea how to actually use them? In this workshop, we will take a beginner-to-expert functional walkthrough of various applications of NLP techniques through an interactive Jupyter Notebook. In particular, we will have a focus on NLP Semantic Embeddings and their various applications in Search, Content/User Understanding, Topic Modelling and Semantic Clustering. After this workshop, the provided Notebook will have direct drop in code examples, both with open source and API hosted solutions, that can be used right away in your projects.

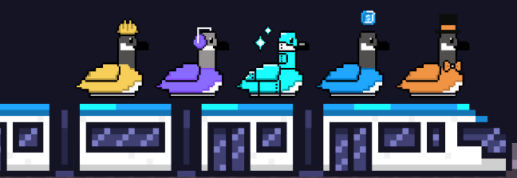
Learning Outcomes

After this workshop, you will be able to:

- Create search models to retrieve relevant multilingual text documents/items and augment them with real world knowledge, and understand how they work
- Classify text into custom categories, and how they can be used for model user intent
- Visually represent and mine massive amounts of documents/text into topics and extract insights from them.

Prerequisite Knowledge

- Python
- Jupyter Notebooks/Google Collaboratory
- (optional) Basic Linear Algebra
- (optional) Basic Machine Learning



Pre-Workshop Checklist

Before the workshop, please make sure you complete the following items:

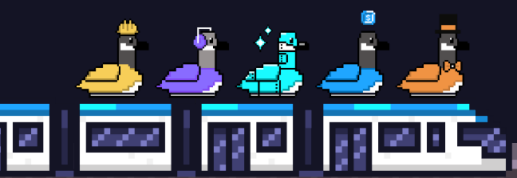
- Create a Google Account and become familiar with Google Colaboratory
- (optional but encouraged) Refresh your knowledge on linear algebra vectors (what they are) and vector distance metrics (Cosine, Euclidean Distance)
- (optional) Create a Cohere account and create an API key

Technical Jargon and Definitions

- (Semantic) **Embedding**: A series of numbers (vector) which represent the semantic meaning of content/text, typically consisting of 768 numbers or more. Similar embeddings have similar meaning.
- **NLP**: Natural Language Processing, a family of AI techniques which focus on unstructured text. In particular, text which is spoken naturally.
- **LLM:** Large Language Model, a large AI model that is trained on Trillions of documents that can extract meaning from text and (optionally) generate new text/continuations.
- **Document:** A collection of text. Documents can be small (a sentence) or large (pages).
- **Topic Modeling:** Methods which seek to group large amounts of documents/text into categories, without knowing ahead of time what those categories are (unsupervised).
- **Semantic Clustering/Classification:** Similar to topic modeling, except categories are predefined with pre-existing data. Here, the task becomes how to assign unknown data to defined categories by inspecting examples of each category.

Timeline (1 Hour)

Time	Module	Description
5 mins	What is Natural Language Processing and what are Semantic Embeddings?	Definitions, use-cases
25 mins	Document Search and Entity Retrieval	How to build your own search engine to retrieve relevant information, and how to enhance text with real world information (wikipedia)



15 mins	Topic Modeling and Semantic Clustering	How to generate/retrieve topics in large amounts of text, and how to label new data into topics/groups
15 mins	Content Understanding and Classification	How to create/use classifiers to extract insights from individual text

Workshop Lead Contact

Liam Hebert
liam.hebert@uwaterloo.ca https://www.linkedin.com/in/liamhebert/

Additional Resources

Hack the North Resources

Hack the North 2023 Event Schedule

Check this out to stay up-to-date on activities, workshops, and other key happenings this weekend.

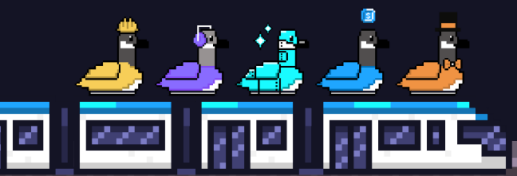
Workshop-Specific Resources

Google Colab Notebook

This notebook will contain all the code examples in this workshop. Ideally, the techniques covered should be able to be copy pasted into your existing projects

HuggingFace Transformers

Industry-standard library for downloading and utilizing pre-trained Large Language models and Computer Vision models.



HuggingFace Transformers

Industry-standard library for downloading and utilizing pre-trained Large Language models and Computer Vision models.

BERTopic

Very powerful and simple pipeline for semantic text topic modeling

Cohere

Easy to use hosted API solution for powerful semantic embeddings. Can be used across multiple languages through a hosted endpoint and offers a free version.