Embeducation

Teacher Guide

This guide outlines the objectives, proposed agenda, materials required, and evaluation methods for the Embeducation workshop. This is followed by a breakdown of and guiding questions/thoughts and answers for each activity (introduction and Modules 0-4), and then additional resources.

Summary

Embeducation is a web application that introduces the idea of how words can be represented by features. It visualizes words by generating word embeddings and plotting these in three-dimensional space. Using the visualization, students can explore nearest neighbors of words, paths between words, and analogies in the 3-D space. Specifically, the application explores word embeddings generated from *Harry Potter: Prisoner of Azkaban*. Students can also explore word embeddings from their own text inputs. The teacher will explain the basic idea behind how word embeddings are generated and guide students through the activity worksheet. The lesson will conclude by asking students to consider how word embeddings are used in practical, real-life applications.

Time Allotment: 80 min

Objectives

- Understand at high level how word embeddings are generated
- Explore how word embeddings encode semantic relationships between words
- Extend concepts of word embeddings to real-life applications

Agenda

- Introduction (10 min)
- Module 0: Pre-Assessment Mind Map (5 min)
- Module 1: Getting Familiar with Embeddings! (20 min)
- Module 2: The Magic of Embeddings with Harry Potter (30 min)
- Module 3: Class Discussion (10 min)
- Module 4: Post-Assessment Mind Map (5 min)

Materials

- 1 laptop per 2 students
- Student Guide
    - Pre-Assessment Mind Map (Module 0)
    - Guided Activities Worksheet (Modules 1 - 3)
    - Post-Assessment Mind Map (Module 4)
- Embeducation Web Application

Assessment

Demonstration of Learning

- Students can explain why word embeddings are important in real-life NLP applications

Classroom Discussion

- Students understand the advantages and shortcomings of word embeddings in both *Harry Potter* and their own text input
- Students can explain how word embeddings relate to current NLP products

Written Responses

- Students complete pre- and post-assessments
- Demonstrate better understanding in post-assessment

Introduction (10 minutes)

Teacher begins by asking “How do computers understand what words mean?” This is natural language processing (branch of artificial intelligence that helps computers understand and interpret human language).

Ask students what might be challenges for computers in understanding. Write these ideas on whiteboard.

Answers to look for include:

- Understanding meaning of words
- How similar two words are
- Computer can be taught things that are wrong (bias in meaning of words)

Module 0: Pre-Assessment Mind Map (5 min)

Have students fill out the mind map. It is okay if students do not have the following answers, but good answers to look for include that Alexa understands how to:

- Split phrases into words
- Understand how words in a phrase relate to each other (“Alexa, remind me to eat sandwiches at noon” is different from “Alexa, eat sandwiches at noon and remind me”
- Differentiate parts of speech (verbs, noun, adjectives, etc.)
- Understand synonyms and words close to each other (“It is a *nice* time to be outside”and “It is a *good* time to be outside” should be interpreted as very close in meaning)

Module 1: Getting Familiar with Embeddings! (20 min)

While students are going through the worksheet, go around and check in on the light bulb questions 💡 and encourage discussion amongst the pairs. Here is an answer key for Module 1:

1.1: Number Line: Answer Key

Expected order: 1, 2, 5, 7, 10 (or reverse)

1.2: Word Line: Answer Key

Expected order: daughter, mother, grandmother, great-grandmother (or reverse)

💡 How did you order the words? *By relative age*

Expected order for additions: son, father, grandfather, great-grandfather (or reverse); should be next to female counterparts

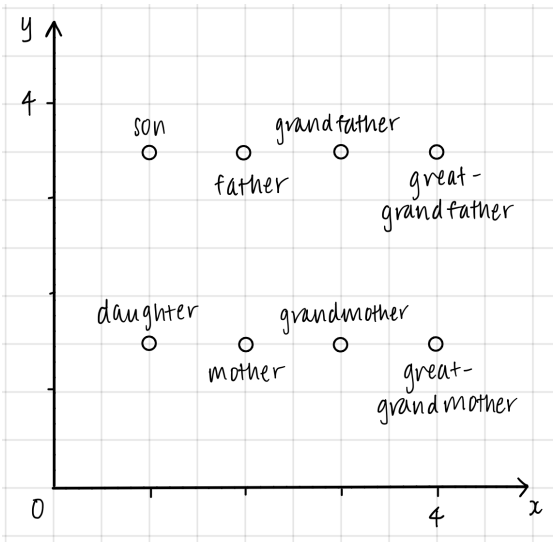
💡 Where did these words go in comparison to the previous set? *Next to female counterparts*

💡 What was harder to to put on the line: the numbers or the words? *Numbers, because there is a strict ordering between numbers, so it’s easy to relate that in space.*

1.3: Word Graph: Answer Key

Check: What is a two-dimensional space? *A geometry where two values are required to determine the position of an element (one value in each dimension)*

Potential graph:



What are the two dimensions you used? *Relative age (x-axis above) and gender (y-axis above)*

1.4: Axis Labeling: Answer Key

x-axis label: size

y-axis label: animal class

1.5: Neighbors: Answer Key

Two nearest neighbors: gorilla, shark

How are these two neighbors related to “whale”? *Closer in size and in animal class than other points*

1.6: Relationships in Space: Analogies: Answer Key

“anchovies” is to “salmon” as “dog” is to “gorilla”

1.7: Paths: Answer Key

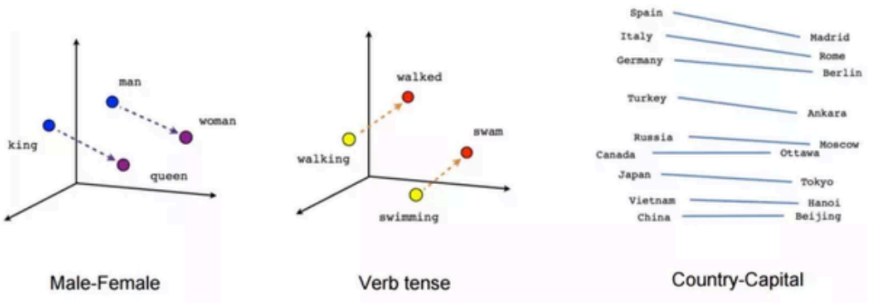
Path goes through “shark”, so final path is: “salmon” ➜ “shark” ➜ “whale”.

- What does this path mean? *The path is a way to get from one word to another in the embedding space. The path should step through related words that get you from salmon to whale.*
- If the path crossed through any points, how are these points related to “salmon” and/or “Whale”? *Shark is a fish that is bigger than salmon, then we take another step in both the size and animal class dimensions, which leads us to whale!*

After students complete worksheet, ask questions about how students were able to discern the relationships between words.

Teacher introduces what word embeddings are and how they help the computer understand meaning in a particular context. Key Ideas:

- Word embeddings convert words to vectors. These vectors capture “feature information” that distinguish each word and some information about what they mean
- Embeddings preserve *semantic relationships* (how words are related based on their meanings)
- Vector Space can be extremely high-dimensional. We used methods called *PCA* or *tSNE* to bring them down to three dimensions that we can easily interpret and visualize
- We can plot these 3D points and explore relationships between words
- Use example of King, Man, Woman, Queen: These are analogies that we see are similarly spaced apart in the 3D space



Module 2: The Magic of Embeddings with Harry Potter (30 min)

2.2: Nearest Neighbors Tool: Answer Key

1. Harry: mcgonagall, malfoy, dudley, hagrid, hippogriff
2. Hermione: ron, weasley, hagrid, ginny, neville
3. Voldemort: wormtail, hedwig, dumbledore, percy, snape
4. Magic: potter, cho, lily, petunia, vernon

Nearest Neighbors Tool: Check for Understanding

- What did students find for each character?
- Why do/don't the neighbors of each character make sense?

Embeddings are not perfect. Since *Harry Potter* is a large series, the context may be slightly different in the different chapters/books. Therefore, the relationships that we expect to see may not be the exact output. There may be nuances in the text that may affect results and not match exactly what we expect to see.

2.3: Paths Tool: Answer Key

1. Ron to Harry: ron ----> hermione ----> hagrid ----> malfoy ----> mcgonagall ----> harry
2. Sirius to Black: snape ----> dumbledore

Paths Tool: Check for Understanding

- What do the these paths from A to B mean?

  Intuitively, a path should step through related words that get you from word A to word B.

- Given a path, what can you infer about the relationship between the first and last word?

  We can see how similar/closely related they are in the context. We expect longer paths to indicate that words are not semantically close in the context in which they are embedded.

2.4: Distance Tool: Answer Key

1. Harry and Dumbledore: 3.359
2. Harry and Voldemort: 7.376
3. Hermione and Magic: 4.450

Distance Tool: Check for Understanding

- Which is larger? Does this make sense?

  It makes sense that Dumbledore is closer to Harry than Voldemort since Harry and Dumbledore are good and Voldemort is bad.

2.5: Analogies: Answer Key

1. Snape to Voldemort is the same as Harry to: Dumbledore
2. James to Lily is the same as Harry to: Cho

Analogies: Check for Understanding

- Are the results what you expected?
- What do these analogies tell you about the distances between these words?

  Both pairs of words should be more or less the same distances from each other. This means that one pair of words is as similar in meaning as the other pair of words are.

2.6: Add Your Own Text: Answer Key

- 5 Nearest Neighbors to 'he' are she, golf, and, likes, to and the 5 nearest neighbors to 'she' are he, to, and, likes, bake
- This could be problematic if, for example, Amazon were providing recommendations based on the 5 nearest neighbours (they don't actually). A female or 'she' might only be recommended things related to baking, even though she loves golf. While a male or 'he' might only be recommended things related to golf, although he loves to bake. This is an example of **bias** in data.

Add Your Own Text: Check for Understanding

- Did you try any other texts?
- Are the results what you expected? Why/why not this be the case?

  These embeddings do not take into consideration the context of the input text. Rather, they are embedded in the context of a large corpus of texts. Therefore, the relationships between words can be thought of as general relationships of the words and how they are in everyday natural human speech/writing.

Module 3: Class Discussion (10 min)

- Why do you think word embeddings are important?

  They allow us to find numerically represent words in a way that computers and algorithms can understand. The representation preserves meaning, so we can obtain meaningful results by adding and subtracting feature vectors (feature representations). We can attempt to teach the computer to understand words like humans do. We can use these feature vectors to train things like neural networks and other machine learning models as well.

- What products do you know of that may use word embeddings?

  Amazon Alexa, Google Home, Spotify, Netflix, Any application that takes natural language as text or speech to give an output

- How might these applications use word embeddings to generate music or movie recommendations?
  - **Music/Video Recommendation Systems:** Need to know what songs are frequently together in very similar contexts
    - Certain words from song/movie titles may be semantically related to a certain genre/mood
    - Using these, we can find nearest neighbors or perform other operations to generate suggestions or find similar content
    - Examples: Amazon Alexa, Google Home, Spotify, Netflix
  - **Sentiment Analysis**: Understand the context under which certain opinions and ratings are generated
    - Pretend you are selling a product on Amazon. You would like to know how well customers like your product. Consider the text corpus of written reviews for your product. For a given embedding scheme and visualization, we can find that all of the positive words cluster together and negative words cluster together. We can see which cluster a particular customer is closer to and which cluster has more customers closer to it.
- Can word embeddings be biased (can we alter how words are related)?
  - Consider the previous example of “She likes to bake and he likes to golf.”
  - We found that “she” is closer to “bake” and “he” is closer to “golf”. We can consider the embeddings to contain some bias towards relating females and

baking more closely than females and golfing (and males and golfing more closely than males and baking).
    - The original text in which words are found may have biases that cause this to happen when we find the embeddings.
- How can we use word embeddings to find potential bias in text?
    - By embedding and visualizing words from a text, we can check if certain words tend to be closer (e.g. “she” and “bake”) than others. We can determine whether this closeness makes sense.
    - We can consider a large corpus of text from media sources to identify any common biases in articles, editorials, etc.
- What did students find exciting/interesting about the whole lesson?
    - Please make note of what students say.
- What was most challenging/difficult to understand?
    - Please make note of what students say.

Module 4: Post-Assessment Mind Map (5 min)

Have your students fill out the mind map to reflect on what they’ve learned. Collect the student worksheets, and observe differences between the two mind maps. Good answers to look for include that Alexa is able to:

- Split phrases into words
- Uses a numeric representation to understand these words and complete its tasks
- Understand how words in a phrase relate to each other (“The happy blue puppy and the sad yellow cat” is different from “The happy yellow puppy and the sad blue cat”). In the former, we expect “blue” to be more closely related to “happy” and in the latter we expect “blue” to be more closely related to “sad”.
- Understand synonyms and words close to each other (“It is a *nice* time to be outside”and “It is a *good* time to be outside” should be interpreted as very close in meaning).

Additional Resources

Tensorflow Guide on Embeddings

Tensorflow Embeddings Projector

Word2vec Algorithm (for producing embeddings)

Snap! Embeddings Activity

More Technical Details of Embeddings