Cooking Up Recipes:

Recipe Recommendation Based on Text Similarity

CS 263

Ashley Chiu
Ritvik Kharkar
Adam Rohde

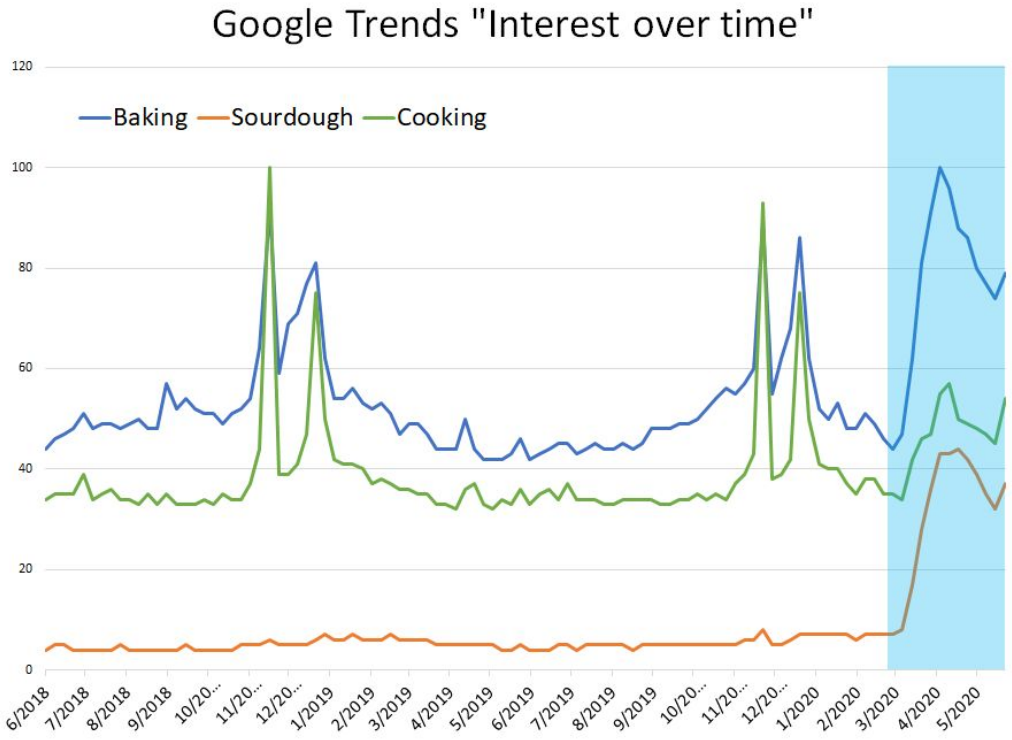
University of California, Los Angeles
Department of Statistics

Motivation

- Inspired by the COVID-19 Shelter in Place Orders and increased necessity to cook at home
- Google trends shows a dramatic surge for cooking and baking related searches coinciding with Shelter-in-Place

How can ***natural language processing help individuals*** embrace this need and trend?

- Use recipe text as a basis for a recipe recommender system
- Such tools like this have need and popularity: commercially developed app ***Yummly***



Dissecting a Cooking Recipe

Cooking recipe: A “structured” text document with unstructured information

- Title
- Ingredient List (with quantity and measurement)
- Sequence of Instructions

Immediate NLP Tasks

- *Document Classification*
- *Word Segmentation and Representation*
- *Semantics*
  - Co-reference Issues
  - Semantic properties of words
  - Relationship extraction
- *Text Similarity/Comparison*

Ultimate Chocolate Chip Cookies

★★★★½

Prep	Total	Servings
15 MIN	1 HR 30 MIN	48

Ingredients

- 2 1/4 cups Gold Medal™ all-purpose flour
- 1 teaspoon baking soda
- 1/2 teaspoon salt
- 1 cup butter, softened
- 3/4 cup granulated sugar
- 3/4 cup packed brown sugar
- 1 egg
- 1 teaspoon vanilla
- 2 cups semisweet chocolate chips
- 1 cup coarsely chopped nuts, if desired

Steps

1. Heat oven to 375°F. In small bowl, mix flour, baking soda and salt; set aside.
2. In large bowl, beat softened butter and sugars with electric mixer on medium speed, or mix with spoon about 1 minute or until fluffy, scraping side of bowl occasionally.
3. Beat in egg and vanilla until smooth. Stir in flour mixture just until blended (dough will be stiff). Stir in chocolate chips and nuts.
4. Onto ungreased cookie sheets, drop dough by rounded tablespoonfuls 2 inches apart.
5. Bake 8 to 10 minutes or until light brown (centers will be soft). Cool 2 minutes; remove from cookie sheet to cooling rack. Cool completely, about 30 minutes. Store covered in airtight container.

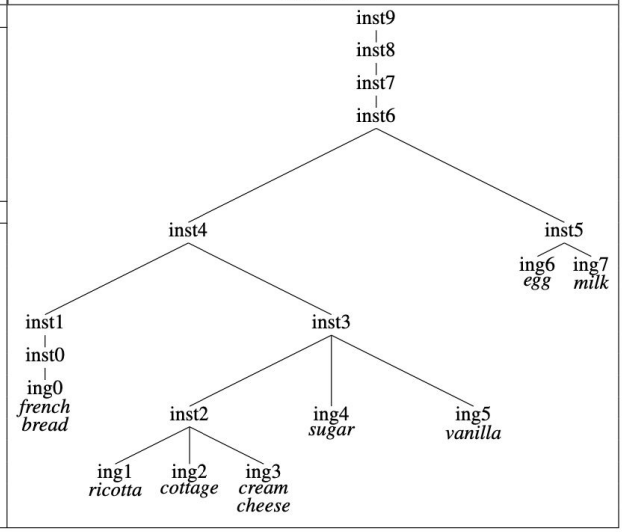
Source: Betty Crocker

Related Works

Predicting the Structure of Cooking Recipes

- Jermsurawong and Habash (2015)
- *SIMMR*:
  Simplified Ingredient Merging Map in Recipes
- Examines high-level flow of ingredients (without modeling the semantics in each individual instruction)

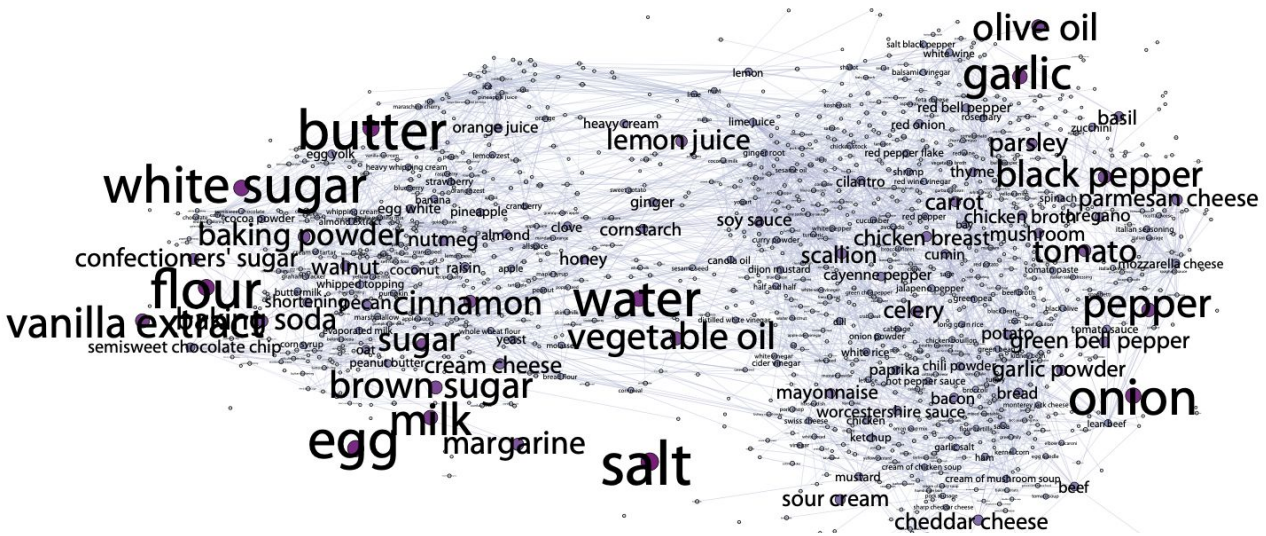
Text Recipe	
ingredients	
0	6 (2 inch thick) slices French bread
1	1/4 cup ricotta cheese
2	1/4 cup cottage cheese, whipped
3	2 tablespoons lowfat cream cheese
4	2 teaspoons white sugar
5	2 teaspoons vanilla extract
6	3 cups egg substitute
7	1/4 cup evaporated milk
instructions	
0	Cut a pocket in each slice of bread.
1	Open carefully
2	In a large bowl, combine the ricotta, cottage cheese and cream cheese.
3	Add the sugar and flavoring extract and beat until smooth.
4	Spread the mixture evenly into each bread pocket.
5	Beat together the egg substitutes and milk.
6	Dip the slices of bread in the egg mixture.
7	Heat a nonstick pan over medium-high heat.
8	Coat with cooking spray.
9	Cook the toast on each side for about 3 to 4 minutes per side until golden brown.



Source: Jermsurawong and Habash (2015)

Recipe recommendation using Ingredient Networks

- Teng, Lin, Adamic (2012)
- Ingredient Complement Network and Ingredient Substitute Network
- Examine co-occurrences and utilize networks to capture underlying “communities” (flavor, functional equivalence, user preference)



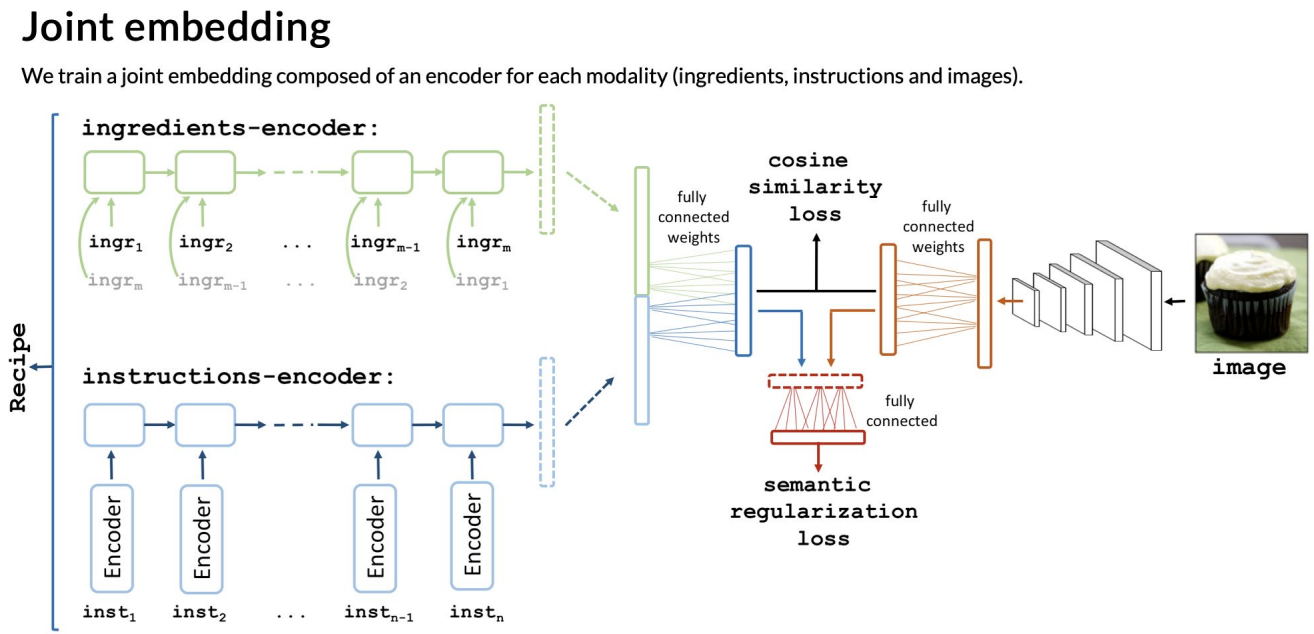
Source: Teng et al. (2012)

Understanding ***Relationships*** between Ingredients and/or Actions

Related Works

Recipe1M+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images

- Marín et al. (2019)
- Develops suitable representations for each of component of a cooking recipe using bi-directional and two-stage LSTMs
- Utilizes cosine similarity to analyze quality of embeddings and rank relevant recipes



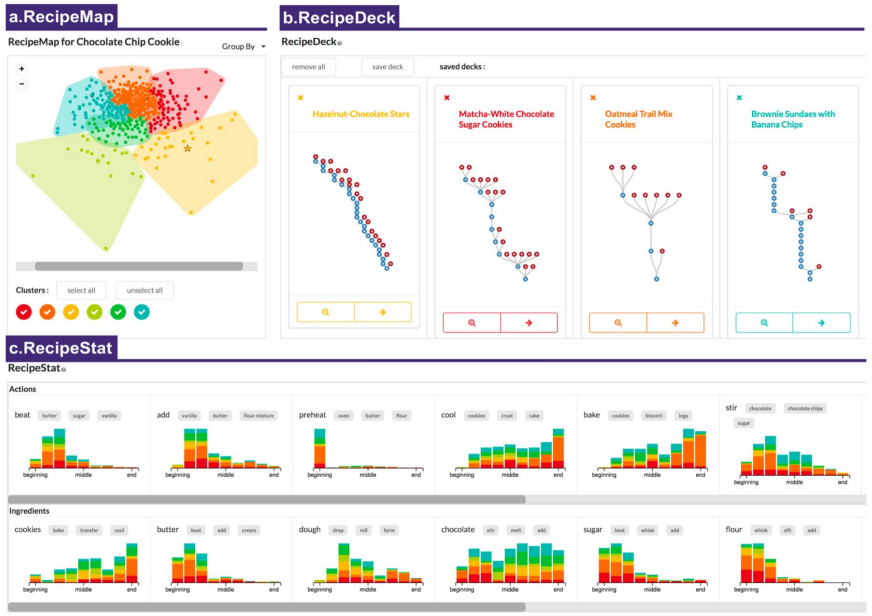
Source: Marín et al. (2019)

Foundations for Text Comparison

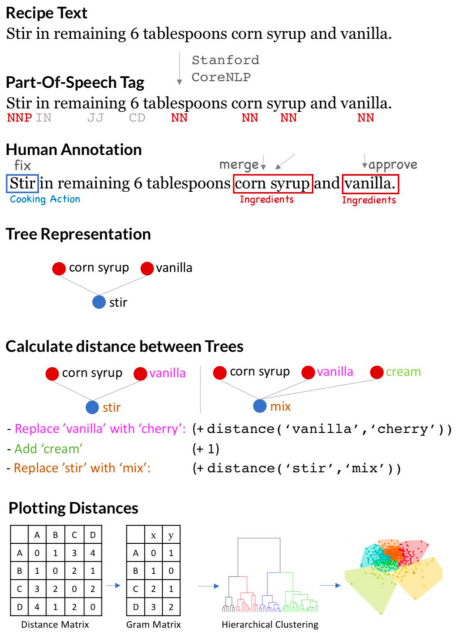
Related Works

RecipeScape: An Interactive Tool for Analyzing Cooking Instructions at Scale

- Chang et al. (2018)
- An interactive system for browsing and analyzing the hundreds of recipes of a single dish
- Developed a computational pipeline that extracts cooking processes from recipe text and calculates a procedural similarity between them, then subsequently clusters recipes into distinct approaches, which capture notable usage patterns of ingredients and cooking actions



Source: Chang et al. (2018)



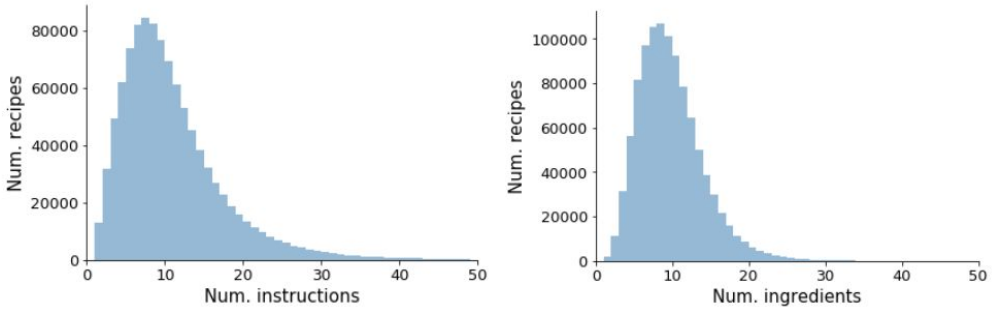
Source: Chang et al. (2018)

Recipe **Classification** and User **Interface**

Recipe 1M+ Data

- Publicly available from MIT CSAIL
- Contains 1M+ cooking recipes from popular websites
- Minimal repeated recipes based (pre-filtered by dataset creators)
- Challenges
  - Volume and size – limited computing power
  - Long tailed vocabulary
  - More on this later
- **Limited our analysis to baking recipes**: must contain the keyboard “bake”
  - Approximately 130K observations

Full Dataset Statistics



Source: Marín (2019)

Example Data

```
['1 (1.5 ounce) bar milk chocolate
candy bar', '1 tablespoon milk', '1
1/2 cups plain fat-free Greek yogurt']

['Put the chocolate bar into a
microwave-safe bowl and add the
milk.', 'Microwave until the chocolate
is soft and melted, about 30
seconds.', 'Stir Greek yogurt into the
chocolate mixture until smooth.']
```

Methods

Recipe 1M+ Data

Learn Vector Representations of Recipes

Multiple approaches to better understand which structural components of recipes are most useful

Bag of Words	Doc2vec	LSTM	BERT
Generate feature vectors	Predict next word in many contexts from recipe using both word & recipe vectors in multiclass classification	Bidirectional LSTM RNN Autoencoder	Learn words by looking at forward and backward sequences *simultaneously*

Generate Cosine Similarity Matrices & Evaluate Quality of Embeddings

Evaluation of Quality of Embeddings:

- Review example recommended recipes
- PCA
- Cluster Analysis

Takeaways:

- Some clustering based on sweet vs savory
- PCs explain fair amount of variation in embeddings

Recommend Recipes!

Highlights

Data Challenges

- **Difficult instructions**
  - “mix wet and dry ingredients”
  - Long instructions
- No meaningful **ingredient order**
- **Ambiguity** of various types
  - “mix” vs “fold”, “ounce” vs “oz”
- **Unlabeled data**
- Large corpus → long tailed vocabulary
- Long run times & limited memory leading to compromises

Recommendation Quality

- **BoW** does surprisingly well
- **Doc2vec** gives reasonable recommendations based on titles
- **LSTM** also does well but with somewhat more variation
- **BERT** is more variable than expected, especially when looking at actual recipe URLs

Results – Post Hoc Quality Evaluation using K-Means

We evaluate the **learned embeddings** and features by running **K-Means** Clustering, then

- Generate a list of most frequent words/ingredients within each of the *K* clusters and
- Identify unique words within each cluster

Bag of Words

Cluster 1:
{'machin', 'fast', 'abm', 'easi', 'unbleach', 'soft', 'vital', 'bagel', 'tabl', 'instant', 'activ', 'extra', 'flax', 'oregano', 'french', 'pita', 'gluten', 'italian', 'sesam', 'hot', 'rye', 'sunflow', 'basil', 'rise', 'warm', 'basic', 'fluffi', 'loaf', 'molass', 'yeast', 'yogurt', 'virgin', 'tomato', 'cornmeal', 'flake', 'skim', 'bun', 'pizza', 'rosemari', 'nonfat', 'dough', 'quick', 'beer', 'homemad'}

Cluster 2:
{'lemon', 'syrup', 'peanut', 'mix', 'cream', 'ginger', 'pie', 'blueberri', 'ice', 'biscuit', 'bar', 'cranberri', 'corn', 'granola', 'low', 'sweet', 'bacon', 'chicken', 'cayenn', 'roast', 'almond', 'mustard', 'vanilla', 'pork', 'cocoa', 'sauc', 'cooki', 'soy', 'red', 'fat', 'chip', 'pecan', 'pumpkin', 'extract', 'nut', 'appl', 'pure', 'light', 'cornstarch', 'sour', 'mapl', 'egg', 'nutmeg', 'dark'}

Doc2vec

Cluster 1:
{'machin', 'canola', 'easi', 'unbleach', 'thyme', 'plain', 'instant', 'activ', 'granola', 'flax', 'cayenn', 'sesam', 'parsley', 'sunflow', 'rye', 'honey', 'rise', 'warm', 'soy', 'yeast', 'fashion', 'yogurt', 'old', 'cornmeal', 'flake', 'pizza', 'mapl', 'dough', 'sea'}

Cluster 2:
{'clove', 'shorten', 'oatmeal', 'lemon', 'confectioners', 'mix', 'crust', 'peach', 'pie', 'blueberri', 'ice', 'biscuit', 'bar', 'low', 'pastri', 'heavi', 'beef', 'roast', 'pork', 'cocoa', 'cooki', 'molass', 'pumpkin', 'shortbread', 'appl', 'cook', 'cornstarch', 'sour', 'nutmeg'}

LSTM

Cluster 1:
{'roast', 'plain', 'cranberri', 'instant', 'pizza', 'easi', 'nut', 'granola', 'oatmeal', 'molass', 'blueberri', 'mix', 'pumpkin', 'dough', 'quick', 'bar'}

Cluster 2:
{'substitut', 'machin', 'canola', 'unsweeten', 'cornmeal', 'cocoa', 'carrot', 'soy', 'red', 'low', 'sour', 'parmesan', 'heavi', 'bacon', 'vegan', 'free'}

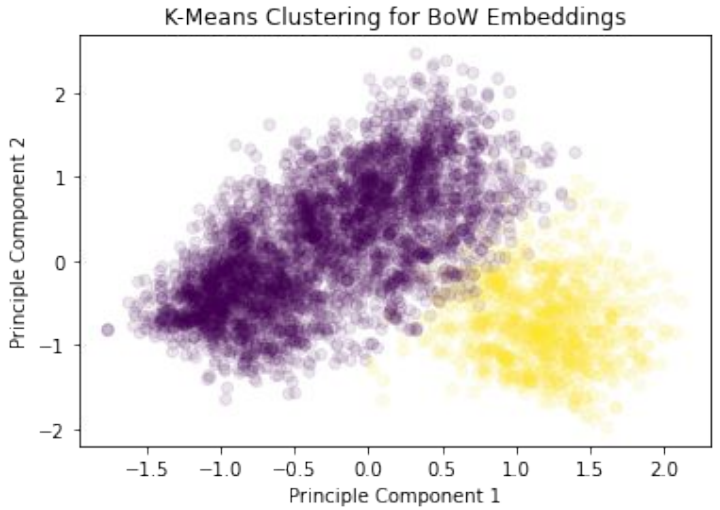
BERT

Cluster 1:
{'machin', 'plain', 'pork', 'cayenn', 'molass', 'mix', 'blueberri', 'sea', 'bar'}

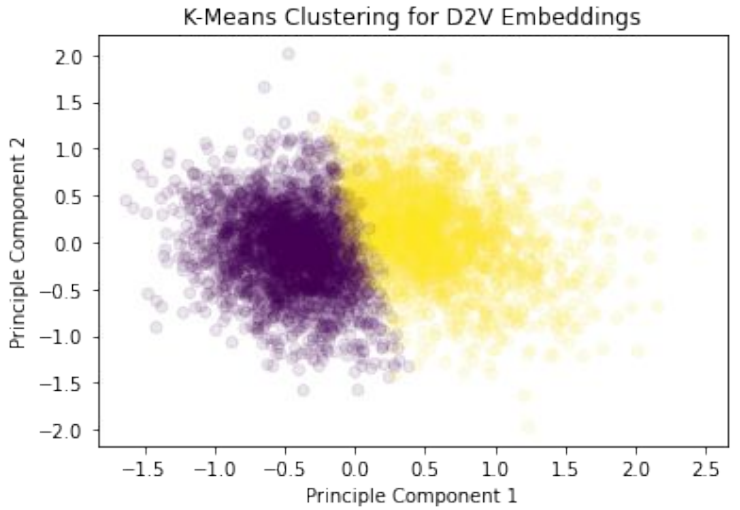
Cluster 2:
{'cornmeal', 'flake', 'canola', 'rise', 'easi', 'soy', 'low', 'bacon', 'ice'}

Results – Post Hoc Quality Evaluation

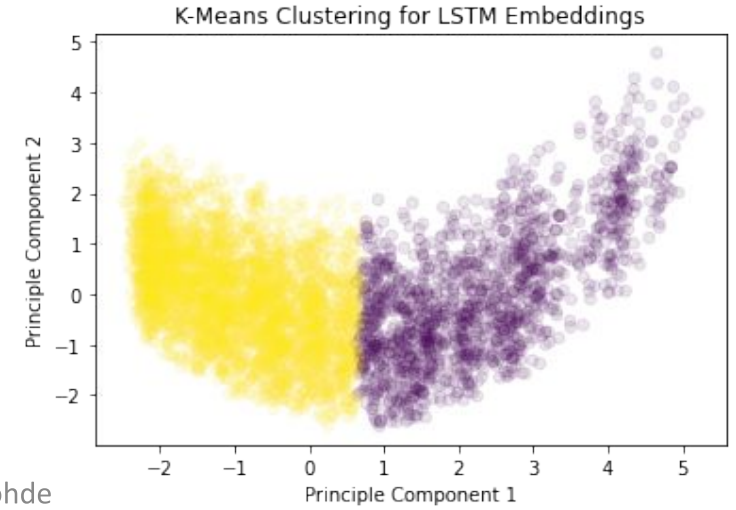
Bag of Words



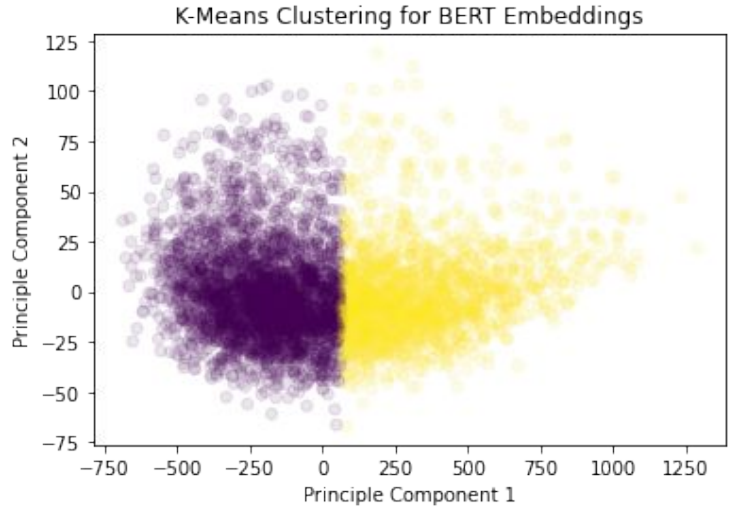
Doc2vec



LSTM



BERT



Concluding Remarks