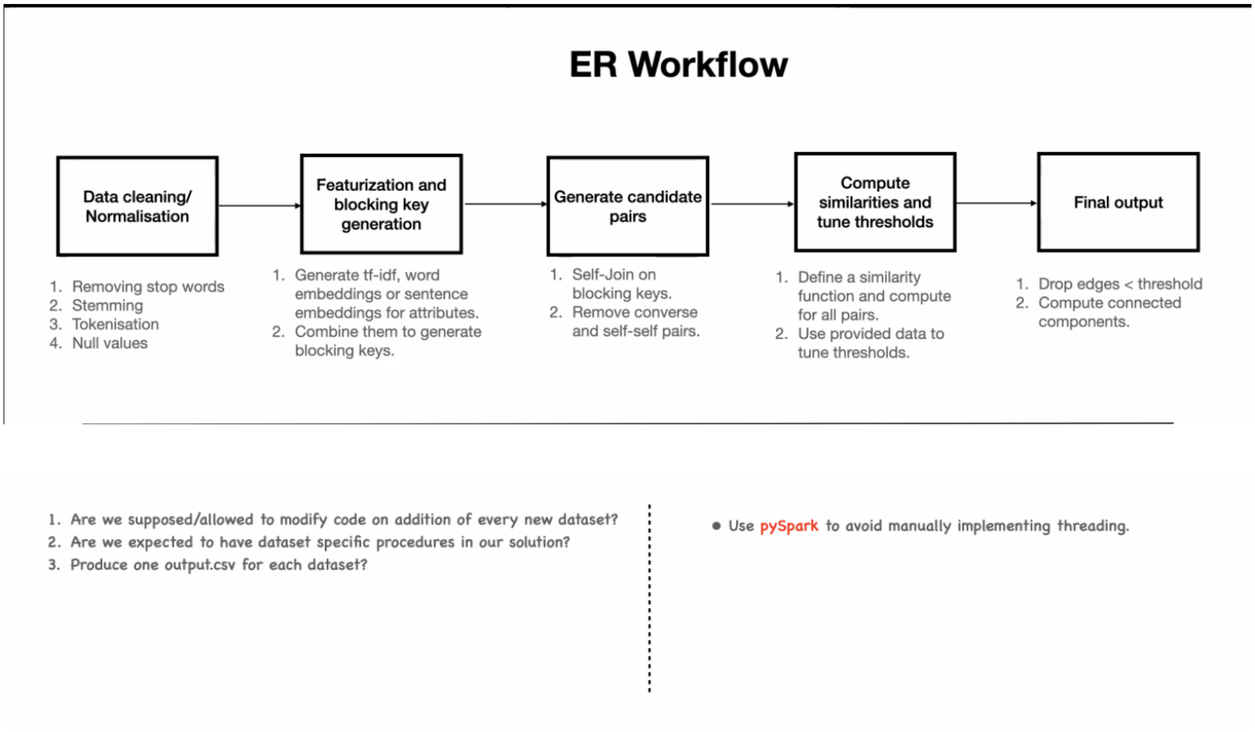
SIGMOD Meeting #1

Class	CS 422
Created	@Mar 17, 2021 5:39 PM
Materials	
Notes	
Reviewed	□
Type	



ER Workflow

Data cleaning/ Normalisation → **Featurization and blocking key generation** → **Generate candidate pairs** → **Compute similarities and tune thresholds** → **Final output**

Data cleaning/ Normalisation

1. Removing stop words
2. Stemming
3. Tokenisation
4. Null values

Featurization and blocking key generation

1. Generate tf-idf, word embeddings or sentence embeddings for attributes.
2. Combine them to generate blocking keys.

Generate candidate pairs

1. Self-Join on blocking keys.
2. Remove converse and self-self pairs.

Compute similarities and tune thresholds

1. Define a similarity function and compute for all pairs.
2. Use provided data to tune thresholds.

Final output

1. Drop edges < threshold
2. Compute connected components.

1. Are we supposed/allowed to modify code on addition of every new dataset?
2. Are we expected to have dataset specific procedures in our solution?
3. Produce one output.csv for each dataset?

- Use pySpark to avoid manually implementing threading.

https://towardsdatascience.com/practical-guide-to-entity-resolution-part-1-f7893402ea7e

Three pathways:

1. Partitioning (similar to k-means / agglomerative clustering)
2. Blocking keys (hash joins)
3. Combinations (ensemble)

Tasks:

- Angelika - Understand the data - explore and identify what it means when rows match (which columns are most similar, etc), compute statistics about how many matching rows, how many missing values, etc. + Snowman
- Akash - Clean the data - stemming, stop words, dealing with missing values
- Vinitra - Data encoding techniques - embedding techniques (sentences - FastText), words (TF-IDF)
- Eleni - Simplest distance techniques - Jaccard, edit-distance + threshold to create initial predictions (baseline accuracy)
    - next step: when there's a null value, don't compute distance w that
- Tell Panos what we're choosing (at this meeting next week)