CS410 Text Information Systems
Course Project – Proposal

Project Title: Text Retrieval Engine for Construction Equipment
Group Name: Kaz Lone Star
Member: Kazuhei Sasaki (ksasaki2@illinois.edu)

Motivation: My employer receives ERP data from major equipment rental companies, aggregates the data, and provides benchmark insights for the industry. One of the biggest challenges in this business model is how to normalize equipment information: the benchmarks are calculated based on client-provided make/model information, but different companies have different representations and conventions of equipment information. In addition, there are countless variations of minor trims in the world of construction equipment. For example, Caterpillar 336 and 336GC are essentially treated as the same model, although there could be a very minor difference. See examples using actual data below.

Provided Make	Provided Model		Normalized Make	Normalized Model
GENIE	77683GT GLAZER		Genie	77683-S
CAT	336GC		Caterpillar	336
CASE IH	ST600	➔	Case	Steiger 600
Challenger	MT743	➔	Agco	Challenger MT743B
Komatsu Construction	PC240LC-11LF	➔	Komatsu	PC240LC-11
CATERPILLAR INC.	D6KLGPAR		Caterpillar	D6K LGP

Examples of Make/Model normalization

Approaches: I am planning to structure this as a document retrieval problem. For the document collection, constructionequipmentguide.com has a fairly large equipment database for a wide range of product categories (e.g. Excavators, Wheel Loaders etc.). I am going to scrape from the webpage and collect over 16,000 make-model combinations. My proposed system accepts a set of make/model as a query, calculates similarities with the document database, and returns Top N matching results depending on the parameter and score. The goal is not to just return the exact match, but rather to return the closest answer from the database. I am currently planning to use n-gram features, but this is to be decided. I may need to implement pre-processing methods to adjust irregular inputs. For evaluation purposes, I plan to use Mean Reciprocal Rank. Additionally, I will develop an interface (either API or a more user-friendly one) that can take a user query and returns the results. If there is time and capacity, I will also think about feedback methodologies. I will primarily use Python for the project. I may use JavaScript and HTML for the front-end work.

Estimated Work Time:

1. Scraping from constructionequipmentguide.com (5h)
2. Build an evaluation dataset, such as hand-labeling (5h)
3. Build prototype retrieval systems on Jupyter Notebook (8h)
4. Build evaluation strategies (2h)
5. Finalize the backend algorithms for the retrieval (5h)
6. Develop API using FastAPI (5h)
7. (Optional) Develop frontend webpage using the API (10h)
8. (Optional) Implement user feedback (10h)

Total: 30-50h