CS410 Text Information Systems
Course Project – Progress report

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

Collecting dataset:

(1) Document dictionary
I was able to scrape 17,464 unique make-model pairs from constructionequipmentguide.com, which is an industry-standard equipment database. See a summary of the data in Appendix 1.

(2) Real ERP data
My employer receives ERP data from major equipment rental companies. I am going to use this actual, anonymized data for performance evaluation. There are 250,652 make-model pairs, manually categorized by our team. I will not be able to publish this dataset due to confidentiality, but I have attached a summary table in Appendix 2.

Building evaluation dataset:

I decided to focus on these 10 categories* and randomly sampled 50 Make-Model pairs for each category from the real ERP data. Note that Generators category is not a part of the document dictionary, but I still included it because (1) I wanted to test cases of no-matching and (2) it is one of the significant product categories to our business.

* Tractors / Excavators / Wheel Loaders / Dozers / Articulating Boom Lifts / Skid Steer Loaders / Forklift Trucks / Generators / Scissor Lifts / Backhoe Loaders

Building prototype retrieval systems:

I have developed a function to return a letter-level n-gram (n=1, 2, 3…) vector and calculate cosine similarity between such vectors. The table below shows some of the samples and corresponding make-model pairs recommended by this approach (i.e. the pair having the highest similarity value).

Category	Make	Model	Proposed Make	Proposed Model	Similarity
Articulating Boom Lifts	GENIE INDUSTRIES	Z34E	Genie	Z-34/22N	0.728869
Backhoe Loaders	CAT	416F2 C4SX	Caterpillar	416F2	0.787186
Dozers	Komatsu Construction	D37PX-21A	Komatsu	D37PX-23	0.422407
Excavators	CASE EXCAVATORS	CX250D LR	Case	CX250D LR	0.479595
Generators	MA	GA-6HB	Bomag	413	0.393051

I am currently facing a few challenges and will continue to work on them:

(1) How to conclude no-match if there is really no matching result? How can I set cutoff?
(2) Are similarity-based approaches effective?
(3) The earlier part of model is generally more important. Can I weigh the n-gram vector?

Developing API:

I do not have much prior knowledge in API development. In an effort to productionize this text retrieval system, I am currently taking a Udemy course about FastAPI. I hope to make this text matching engine available to public via the API.

CS410 Text Information Systems
Course Project – Progress report

Estimated Work Time:

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

Total: 30-50h

Appendix 1. Scraped data from constructionequipmentguide.com, count by equipment category

Equipment Category	Count	Equipment Category	Count
Tractors	3,474	Log Loaders	35
Excavators	2,630	Forwarders	34
Forklifts	1,726	Feller Bunchers	32
Wheel Loaders	1,380	Carrydeck Cranes	32
Smooth Drum Rollers	1,194	Concrete Pavers	30
Crawler Dozers	893	Walk / Tow Behind Compactors	30
Aerial Lifts	722	Wheel Dozers	29
Skid Steer Loaders	400	Combination Rollers	29
Off-Highway Trucks	337	Trench Compactors	24
Backhoe Loaders	330	Landfill Compactors	23
Cab / Chassis Trucks	293	Dumpers	19
Balers	234	Semi Trailers	19
Compact Track Loaders	234	Air Seeders / Carts	15
Motor Graders	233	Pipe / Pole Trailers	15
All Terrain Cranes	229	Light Towers	13
Truck Cranes	218	Utility Trailers	13
Asphalt Pavers	217	Skip Loaders	11
Rough Terrain Cranes	211	Oilfield Trailers	11
Padfoot Compactors	208	Car Carrier / Transport Trailers	11
Combines	193	Pipelayers	8
Crawler Cranes	156	Rakes / Tedders	8
Floaters / Sprayers	147	Rippers Ag	7
Drop Deck / Lowboy Trailers	143	Reel / Cable Trailers	7
Field Cultivators	132	Knuckleboom Cranes	6
Tag Trailers	126	Manure Handling	5
Crawler Loaders	112	Flatbed Trailers	5
Pneumatic Rollers	96	Curb / Gutter Machines	4
Tilt Trailers	94	Rammers	3
Material Handlers	77	Mini Cranes	3
Dump Trailers	68	Bucket Trucks	3
Cold Planers / Milling Machines	64	Mowers	2
Scrapers	60	Agricultural Trailers	2
Skidders / Yarders	60	Plows	1
Tower Cranes	56	Headers	1
Processors / Harvesters	56	Dry Van Trailers	1
Gooseneck Trailers	48	**Total**	**17,464**
Mower Conditioners / Windrowers	43		
Soil Compactors	40		
Traveling Axle Trailers	39		

Appendix 2. Real ERP data, count by equipment category

Equipment Category	Count	Equipment Category	Count
Excavators	17,783	Lighting Equipment	528
Concrete Equipment	14,586	Transport Trucks	514
Earthmoving Attachments	13,535	Other Trailers	502
Forklift Trucks	11,877	Tanks And Boxes	470
HVAC	10,467	Water Trailers	454
Lawn And Landscape	9,995	Hydraulic Tools	450
Generators	9,911	Dump Trailers	393
Pumps	9,199	Fuel, Tank, And Vacuum Trailers	387
Surface Treatment	8,047	Crawler Cranes	379
Air Tools	7,351	Power Equipment	369
Electric Tools	7,155	Site Dumpers	368
Heavy Earthmoving Attachments	6,649	Service Trucks	337
Air Compressors	6,351	Storage Containers	335
Scissor Lifts	6,027	All Terrain Cranes	334
Telehandlers	5,937	Scrapers	325
Light Compaction	5,600	Off-Highway Water Trucks	273
Articulating Boom Lifts	5,281	Vehicles	241
BULK/RE-RENT	5,212	Crawler Loaders	229
Telescopic Boom Lifts	4,717	Pneumatic Rollers	224
Compact Track Loaders	4,384	Air Equipment	206
Dozers	4,383	Truck Cranes	199
Wheel Loaders	4,202	Material Handlers	182
Material Handling	3,947	Soil And Landfill Compactors	168
Other Equipment	3,389	Welding Tools	156
Light Vehicles	3,376	Engines	154
Skid Steer Loaders	3,335	Construction Hoists	123
Warehouse Equipment	3,012	Compact And Mini Cranes	122
Single Drum Rollers	2,896	Off-Highway Haul Trucks	120
Backhoe Loaders	2,856	Tower Cranes	100
Tag-Along Trailers	2,570	Pipelayers	79
Tractors	2,548	Crane Attachments	41
Welders	2,521	Semi Trailers	41
Light Towers	2,420	Aerial Attachments	38
Double Drum Rollers	2,331	Other Cranes	31
Boom Trucks, Bucket Trucks, And Digger Derrick	2,084	Other Aerial	28
Sweepers And Brooms	2,000	Railroad Equipment	26
Other Trucks	1,849	Container Handling Equipment	12
Temporary Assets	1,773	Wheel Dozers	10
Trenching Equipment	1,635	Front Shovels	8
Aggregate Equipment	1,551	Scaffolding And Staging	6
Pickup Trucks	1,451	Ride-On Compaction Attachments	5
Vertical Mast Lifts	1,256	Truck Tractors	5
Paving Equipment	1,207	Tractor Attachments	2
Motor Graders	1,151	Office, Accommodation, And Welfare Units	2
Carry Deck And Pick-And-Carry Cranes	1,033	Aircraft	2
Articulated Dump Trucks	969	Overhead Cranes	1
Mini Dumpers And Loaders	939	**Total**	**250,652**
Forestry Equipment	886		
Ladders	882		
Agricultural Equipment	851		
Track-Driven Equipment	814		
Water Trucks	797		
Dump Trucks	754		
Forklift Accessories	753		
Rough Terrain Forklifts	713		
Rough Terrain Cranes	701		
Towable Boom Lifts	621		
Trench Shoring	594		
Traffic Control	559		