Gcloud operations:

**Push data:**

gsutil cp "C:\Users\Austin Bell\Documents\Columbia\Big Data\project\Data\product\_aspects\_electronics\_and\_games.csv" gs://%BUCKET%/aspect\_ranking/product\_aspects\_electronics\_and\_games.csv

gsutil -o GSUtil:parallel\_composite\_upload\_threshold=150M cp "C:\Users\Austin Bell\Documents\Columbia\Big Data\project\Data\item\_dedup.json" gs://%BUCKET%/aspect\_ranking/item\_dedup.json

gsutil -o GSUtil:parallel\_composite\_upload\_threshold=150M cp "C:\Users\Austin Bell\Documents\Columbia\Big Data\project\Data\metadata.json" gs://%BUCKET%/aspect\_ranking/metadata.json

gsutil cp "C:\Users\Austin Bell\Documents\Columbia\Big Data\project\Data\linking\_verbs.txt" gs://%BUCKET%/aspect\_ranking/linking\_verbs.txt

**Initialize clusters**

gcloud beta dataproc clusters create project --optional-components=ANACONDA,JUPYTER --image-version=1.4 --enable-component-gateway --bucket %BUCKET% --project %PROJECT% --num-workers 2 --worker-machine-type n1-standard-8 --master-machine-type n1-standard-8 --properties spark:spark.jars.packages=JohnSnowLabs:spark-nlp:2.2.1 --metadata "PIP\_PACKAGES=spark-nlp==2.2.1 tqdm spacy” --initialization-actions gs://dataproc-initialization-actions/python/pip-install.sh --region europe-west1

**Run Data Extraction**

gcloud dataproc jobs submit pyspark --cluster project extractData.py --driver-log-levels root=FATAL -- %BUCKET%

**Run Aspect Extraction**

gcloud dataproc jobs submit pyspark AspectExtraction.py --cluster project --region europe-west1 --driver-log-levels root=FATAL -- %BUCKET%

**Run Clustering Job**

gcloud dataproc jobs submit pyspark --cluster project --region europe-west1 --properties=spark.jars.packages=JohnSnowLabs:spark-nlp:2.2.1 --driver-log-levels root=FATAL ClusterAspects.py -- %BUCKET%