Tutorial: Solr for dealing with large scale language model

Tutorial Plan

1. Short Introduction to Solr indexing engine

2. Basic Installation and Configuration Guidelines

3. Howtos for language models:
    a. Ingesting language model data
    b. Updating language models
    c. Querying Solr to retrieve n-grams

Short Introduction to Solr indexing engine

- Solr:
  - Solr is a standalone enterprise search server with a REST-like API.
  - Solr enables you to easily create search engines.
  - Solr uses the Lucene search library and extends it.
  - Apache Lucene is a high performance search & information retrieval technology.

Basic Installation and Configuration Guidelines

- Java 8 required
- Download latest version of solr
  - Solr 7.1.0
- Unzip it, go to the folder
  - Run command ***bin/solr start***
- Open solr interface in browser at port **8983 .**
- Create core
  - bin/solr create -c <core name>
- Solr set an unique document id for each document,which is signed ints, so a single index were limited to 2^31-1
- Can set custom document id
  - Phrase can be doc-id

Basic Installation and Configuration Guidelines (cont.)

- Configuration file reside → **<Solr folder>/server/solr/<core name>/conf**
- Change in two file if you need to set doc id by own
  - managed-schema
  - solrconfig.xml

Ingestion

- Ingestion through .csv file:
  - curl --noproxy <**your_ip**> "http://**your_ip**:8983/solr/**solr_core**/update?commit=true" --data-binary @test_data.csv -H “Content-type:application/csv”
- Ingestion through .json file:
  - curl --noproxy <**your_ip**> "http://**your_ip**:8983/solr/**solr_core**/update?commit=true" --data-binary @test_data.json -H “Content-type:application/json”
- Single document ingestion is possible. For details you can see here

Ingestion & Updation


```
[
{       "phrase": "Online Tutorials",
        "n_gram": 2,
        "occurrence_count": {
        "inc": 10
        } },
  {     "phrase": "science",
        "n_gram": 1,
        "occurrence_count": {
        "inc": 10
        } }
]
```

Query

- Reg-ex query: e.g. **phrase_lowercase:/o.* t.*/**
- Exact match query: e.g. **phrase_lowercase:"online tutorials"**
- Java library: solrJ
- Libraries available in other languages also