LUCENE / SOLR
REVOLUTION 2016

OCTOBER 11-14, 2016 • BOSTON, MA

LUCENE / SOLR REVOLUTION 2016

Autocomplete Multi-Language Search Using Ngram and EDismax Phrase Queries

Ivan Provalov

Sr Software Engineer, Netflix

Overview

- Use Case
- Configuration, scoring
- Language challenges
- Character mapper
- Query testing framework

Going Global at Netflix

- Netflix launched globally in January 2016
- 190 countries
- Currently support 23 languages

NETFLIX Browse DVD planet tester

Explore titles related to: Planet of the Apes | The Planets | Planet 51 | Planet Hulk | Rise of the Planet of the Apes

Use Case

- Video titles, person's names, genre names
- Shorter documents should be ranked higher
- Autocomplete
- Recall over precision for lexical matches (click signal corrects this)

Configuration

- Solr 4.6.1
- Edismax: boosting, simple syntax, max field field score
- Phrase: prevents from cross field search
- Ngram: character ngram search

Character Ngram Search

“Breaking bad”

b - 0

br - 0

bre - 0

brea - 0

break - 0

breaki - 0

breakin - 0

breaking - 0

b - 1

ba - 1

bad - 1

Scoring

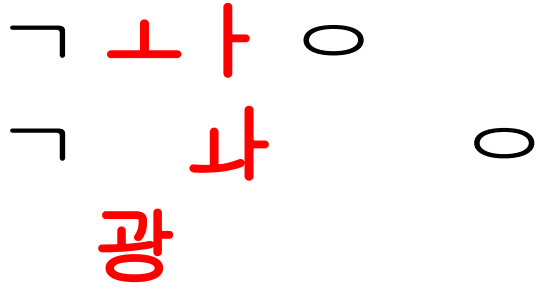
- Skewed data distribution (e.g. one field sparsely populated)
- Doc length normalization
- Unigram language model
- Term Frequency / Terms in Doc
- Log to avoid underflow errors
- Negative score (5.5.2 Dismax Scorer breaks)

Language Challenges

- Multiple Scripts
  - Japanese: Kanji, Hiragana, Katakana, Romaji
- No token delimiters: Japanese, Chinese
- Korean character composition
- Stopwords and autocomplete
- Stemming

Korean: Character Composition

- input jamo
- decomposed jamo
- fully composed hangul



Japanese: Multiple Scripts

- ‘南極物語’ (‘Antarctic Story’)
- Tokenizer: 南極 物語
- Reading form: ナンキョク モノガタリ
- Query in Katakana: ナンキョク
- Query in Hiragana: なんきょく
- Transliteration required

Tokenization Pipelines

- Char Filter: pre-processes input characters
- Tokenizer: breaks data into tokens
- Filters: transform, remove, create new tokens

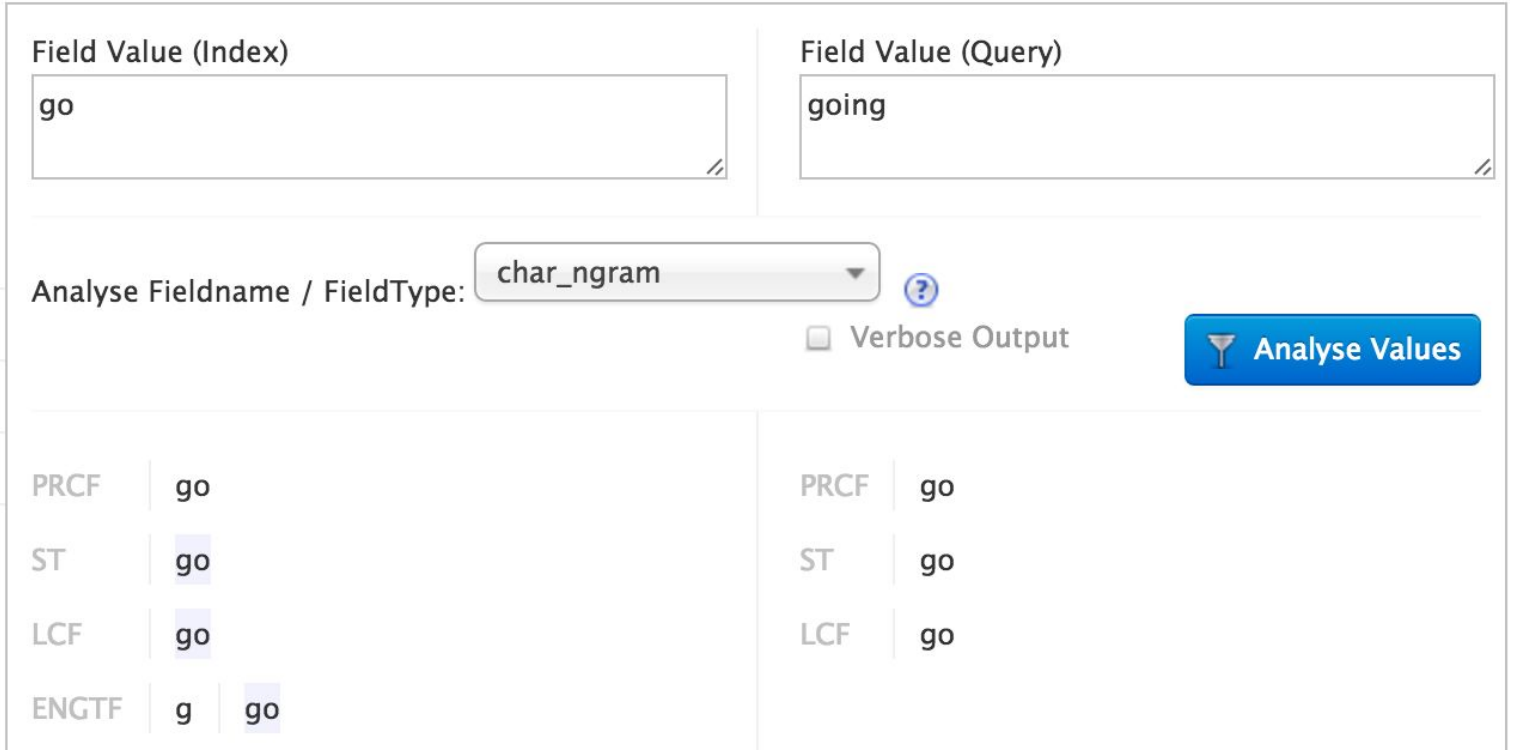
Simple Pipeline Example: index

- CharFilters: PatternReplaceCharFilterFactory
  - pattern: ([a-z]+)ing
- Tokenizer: StandardTokenizerFactory
- Filters: LowerCaseFilterFactory, EdgeNGramFilterFactory

Simple Pipeline Example: query

- CharFilters: PatternReplaceCharFilterFactory
  - pattern: ([a-z]+)ing
- Tokenizer: StandardTokenizerFactory
- Filters: LowerCaseFilterFactory

Simple Pipeline Example



Character Mapping Filter Cases

- Prefix Removal
  - Arabic ال (alef lam)
- Suffix folding
  - Japanese ァ (katakana small a) => ア (a)
- Character decomposition
  - Korean ᅫ (jungseong we) => ᅮ (u) and ᅦ (e)

Character Mapping Filter Cases

- Stemmer implementation, or extension
  - Character mapper reference implementation of the Russian stemmer
- Patch to Lucene
  - LUCENE-7321

Query Testing Framework

- Open source project
- Google Spreadsheets based UI
- Unit tests for languages queries
- Regression testing after changes, upgrades
- 20K queries
- 7K titles

Google Spreadsheets as Input

id	title_en	title_localized	q_regular	q_regular	q_misspelled
1	Fuller House	Huset fullt – igen	Huset fullt	huset	
2	Friends	Vänner	Vänne		Vanner
3	VANish	VANish	van		

Google Spreadsheets as Detail Report

A	B	C	D	E	F
name	failure	query	expected	actual	comments
swedish-video-regular	supersetResultsFailed	van		Vänner	

Diff

name	failure	query	expected	actual	comments
swedish-video-misspelled	noResultsFailed	Vanner	Vänner	NONE	FIXED
swedish-video-regular	supersetResultsFailed	van		Vänner	NEW

Google Spreadsheets as Summary Report

name	titles	queries	supersetResultsFailed	differentResultsFailed	noResultsFailed	successQ	precision	recall	fmeasure	comments
swedish-video-regular	3	4	1	0	0	3	87.50%	100.00%	91.67%	
swedish-video-misspelled	1	1	0	0	0	1	100.00%	100.00%	100.00%	

Diff

name	titles	queries	supersetResultsFailed	differentResultsFailed	noResultsFailed	successQ	precision	recall	fmeasure	comments
swedish-video-regular	0	0	1	0	0	-1	-12.50%	0.00%	-8.33%	
swedish-video-misspelled	0	0	0	0	-1	1	100.00%	100.00%	100.00%	

Summary

- Use case: short fields, autocomplete, P/R
- Configuration, scoring
- Language challenges
- Character Mapper patch (LUCENE-7321)
- Query testing framework
  https://github.com/Netflix/q

References

Query testing framework

Chris Manning IR Book, LM Chapter

Trey Grainger's presentation on Semantic & Multilingual Strategies in Lucene/Solr

Character Mapping Patch and Documentation

Java Internationalization, March 25, 2001, by David Czarnecki, Andy Deitsch