ELASTICSEARCH SUGGESTERS

Beyond Autocomplete

BRETT HAZEN

Principal Consultant at ILM

C#/.NET & JavaScript

Runner

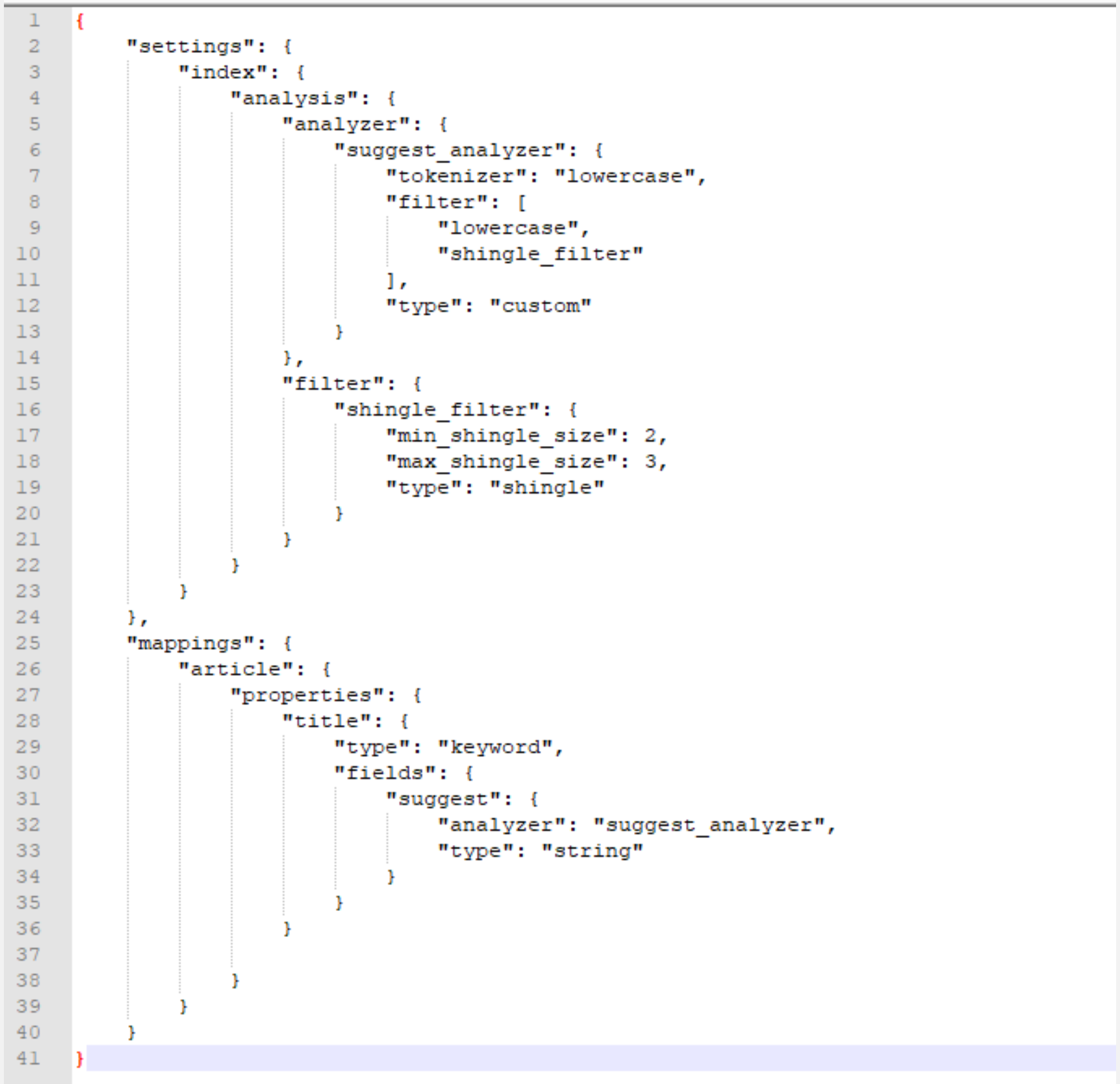
Very Amateur Musician

Bird Owner

WHY THIS TALK?

OUR PLAN

- Cover some Elasticsearch basics
  - Index and mapping
  - Analyzers
- Tour through each type of suggester with example
- Q&A



```
{
    "settings": {
        "index": {
            "analysis": {
                "analyzer": {
                    "suggest_analyzer": {
                        "tokenizer": "lowercase",
                        "filter": [
                            "lowercase",
                            "shingle_filter"
                        ],
                        "type": "custom"
                    }
                },
                "filter": {
                    "shingle_filter": {
                        "min_shingle_size": 2,
                        "max_shingle_size": 3,
                        "type": "shingle"
                    }
                }
            }
        }
    },
    "mappings": {
        "article": {
            "properties": {
                "title": {
                    "type": "keyword",
                    "fields": {
                        "suggest": {
                            "analyzer": "suggest_analyzer",
                            "type": "string"
                        }
                    }
                }

            }
        }
    }
}
```

WHAT DOES AN ANALYZER DO

- Start with the text "Domain Driven Design by Example"
- Standard analyzer
  - No character filters
  - Tokenizer: standard
  - Token filters: standard, lower case, stop
- API for helping debug

WHAT DOES AN ANALYZER DO

- After tokenizer
  - [“Domain”,“Driven”,“Design”,“by”,“Example”]
- After standard filter
  - [“Domain”,“Driven”,“Design”,“by”,“Example”]
- After lowercase filter
  - [“domain”,“driven”,“design”,“by”,“example”]
- After stop filter (actually disabled by default)
  - [“domain”,“driven”,“design”, “by”, “example”]
  - If not disabled
    - [“domain”,“driven”,“design”,“example”]

WHAT IS A SUGGESTER?

- Suggest similar looking terms
  - what that means depends on suggester
- Parts of suggest features still under development
  - documentation doesn’t say which parts
- Three types of suggesters
  - Completion
  - Term
  - Phrase

COMPLETION SUGGESTER

- Provide autocomplete
- Only works based on prefix
- Stored as special data structure for speed
  - Costly to build
  - Stored in memory

TERM SUGGESTER

- Correct a user’s spelling
- Suggestions are based on edit distance

WHAT IS EDIT DISTANCE?

- How many characters need to be changed for a term to match
- Example
    - User provided text is “domian”
    - Corrected term “domain” has an edit distance of 1
        - Only one character had to change for the terms to match

PHRASE SUGGESTER

- Like term suggester, only fancier
- Based on n-gram language model

WHAT IS N-GRAM LANGUAGE MODEL?

- A way of tokenizing a string
- Example
  - We are indexing the string "domain"
  - We are using a gram size of 2
  - We'll get these tokens:
    - do, om, ma, ai, in

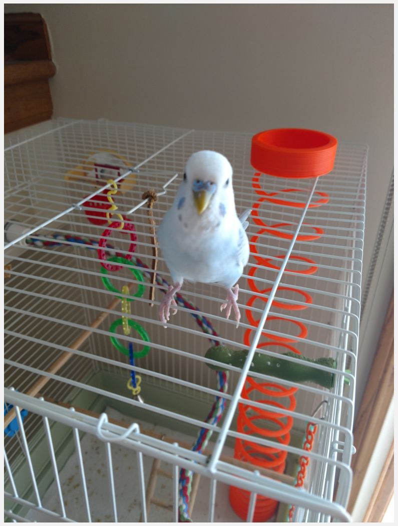
N-GRAM LANGUAGE FOR WORDS

- When we want groups of words instead of groups of characters they're referred to as shingles
- Example
  - "Domain Driven Design"
  - Minimum shingle size of 2
  - Maximum shingle size of 3
  - We would get
    - "Domain Driven", "Driven Design", and "Domain Driven Design"

CONTEXT SUGGESTER

- Used with completion suggester
- Two types of context:
  - Category
  - Geo
- Requires special mapping
  - Increased index size of completion field
    - It's entirely on the heap, tread lightly

QUESTIONS



THANK YOU

- @BrettEHazen
- brett.hazen@ilmservice.com