Using Natural Language Processing to Perform Spatial Searches of Open Street Map Features In ArcGIS

Faculty Advisors: Dr. Alex Klippel
Dr. Jan Wallgrün

Gary Huffman
grh145@psu.edu

Presentation Overview

- Problem Overview
- Search Using ArcGIS Desktop “Out of the box”
- Open Street Map
- Natural Language Processing
- Spatial Language and Spatial Representations
- Proposed System Architecture and Implementation
- Next Steps and Follow-on Work

Problem Overview

Motivation

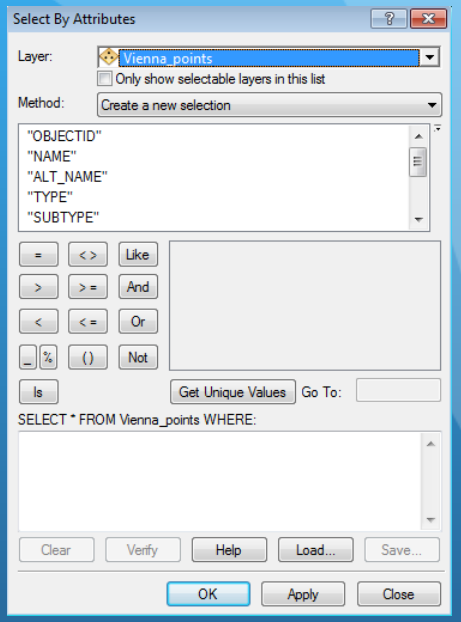
- Limitations of search in ArcGIS Desktop
  - SQL-based -- RDBMS and shapefiles
  - Slow with large databases
- Pervasiveness of Open Street Map
  - Lots of data
  - Contains places of interest that are not available in other data sets
  - Increasing popularity: Apple iPhoto & Four Square
  - Basemap option in ArcGIS Desktop (yet no search)
- User familiarity with natural language search

Objective

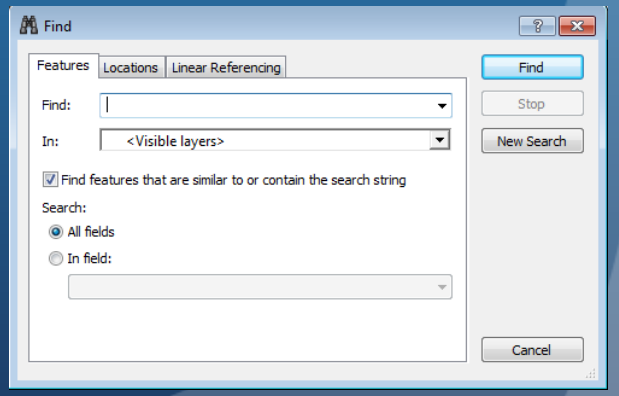
- I will integrate aspects of Natural Language Processing into ArcGIS to search Open Street Map data
  - Spatial Search and Topological Relationships
  - Attribute Search

Search Using ArcGIS Desktop

- Select by Attribute
  - Users construct SQL statements
  - Requires basic understanding of database schema and SQL
  - Use of ‘like’ in query gets close matches

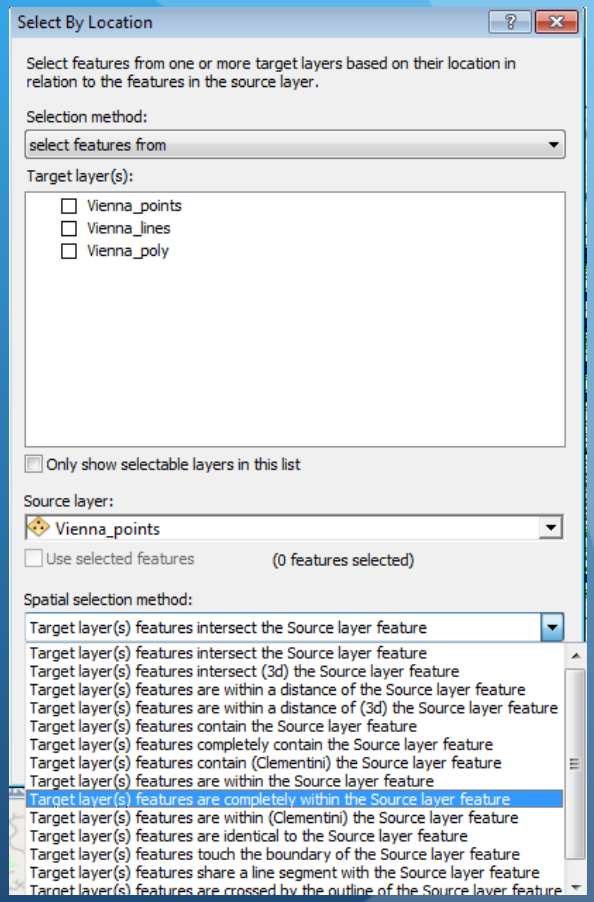


- Find
  - Fuzzy search
  - Inflexible (Starbucks != starbucks)
  - Slow with large datasets



Search Using ArcGIS Desktop

- Select by Location
  - Spatial and topological relationships
    - Containment (In/On)
    - Intersection
    - Equality
    - Nearby (Proximity distance)
  - Specify source and target layers
  - Features selected beforehand
  - Differences and meanings of the spatial selection methods



Open Street Map

- The Wikipedia of geospatial information
- User contributed and moderated data
- Roughly 21GB of compressed XML formatted geospatial data
  - Nodes (Points)
  - Ways (Lines and Polygons)
  - Relations (Lines and Polygons)
- On-line search interface (Nominatim) and a Web Service API
- Available as a basemap layer in ArcGIS Desktop
  - All or nothing
  - Cannot search the basemap
- Available on-line at www.openstreetmap.org

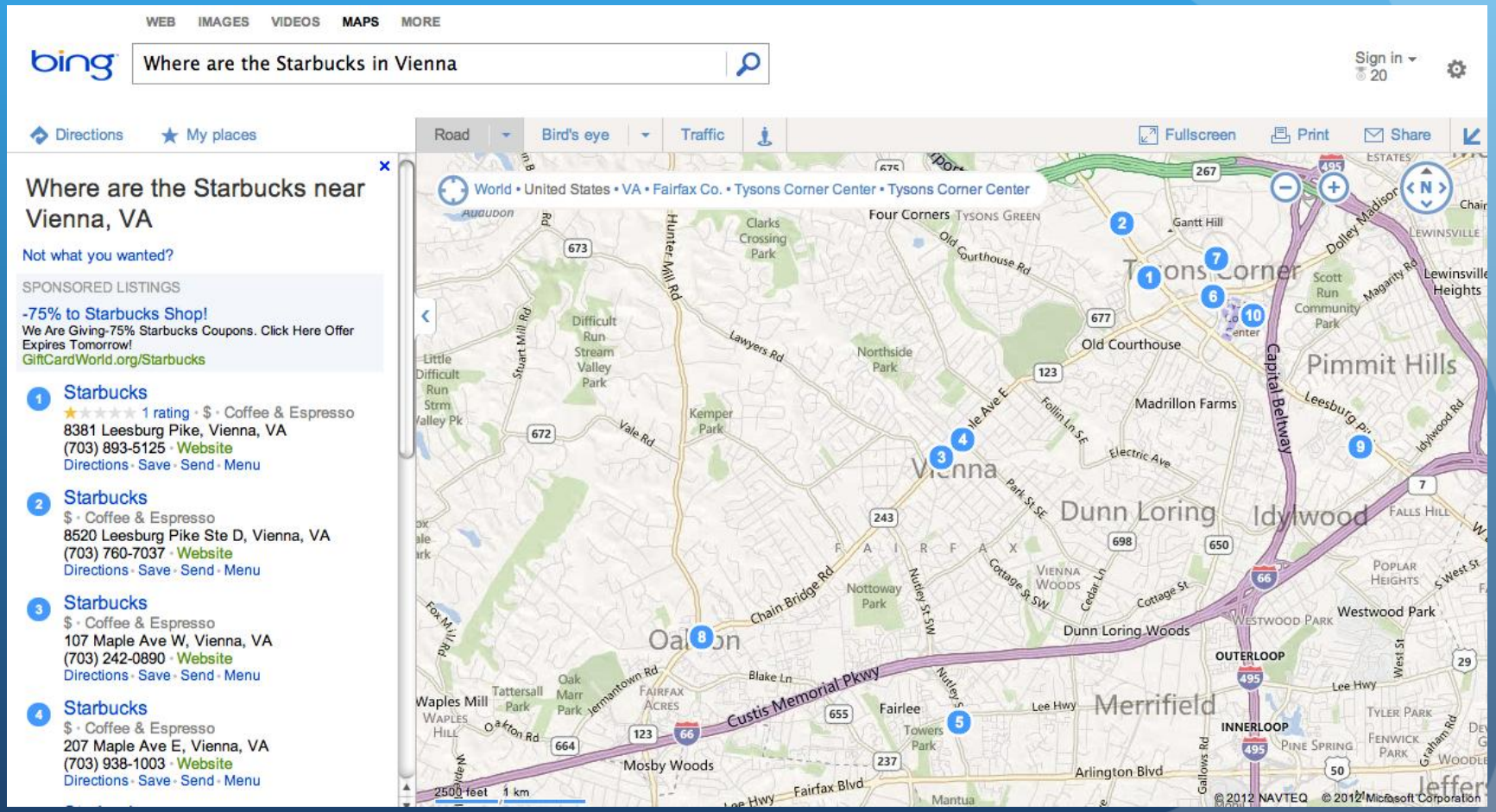
Natural Language Processing

- Natural language - It's how humans talk
  - We say: "Where are the Starbucks in Vienna?"
  - We don't say: "Select * where Name = 'Starbucks' and City = 'Vienna' and State = 'VA'"
- Natural Language Processing
  - Part computer science, part linguistics
  - Goal is to get computers to understand human language
  - Non-trivial problem
    - **Reading**
      - Noun as in "He gave a reading."
      - Verb as in "I was reading earlier today."
      - Proper noun (place name) as in "Reading, Pennsylvania"

Natural Language Processing (cont.)

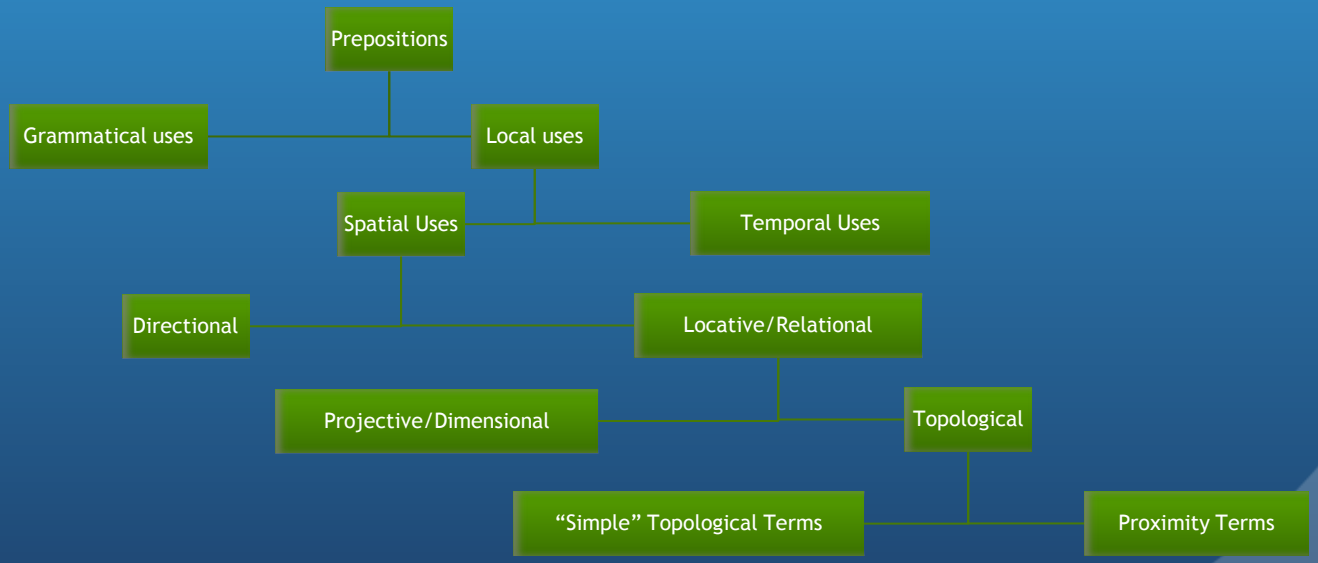
- NLP systems try to understand the linguistic, grammatical and semantic meaning inherent in language
  - Parts of Speech
  - Named Entities
  - Parsing and Tokenization
- Consider the following statements that use the preposition IN:
  - The crack in the jar.
  - The flowers in the vase.
- Systems implementing NLP are all around us – we use them daily
  - Spam/junk e-mail filter
  - Calendar events from e-mail messages
  - Internet search
    - How to repair Maytag dishwasher with leaky door?
  - Internet map searches for geographic information
    - Bing Maps – Where are the Starbucks in Vienna?

On-line Maps and Search



Spatial Language and Spatial Representations

- How do we describe where things are in the world?
  - In language, often through the use of spatial prepositions
  - Where are the Starbucks ***IN*** Vienna?



Coventry and Garrod, 2004

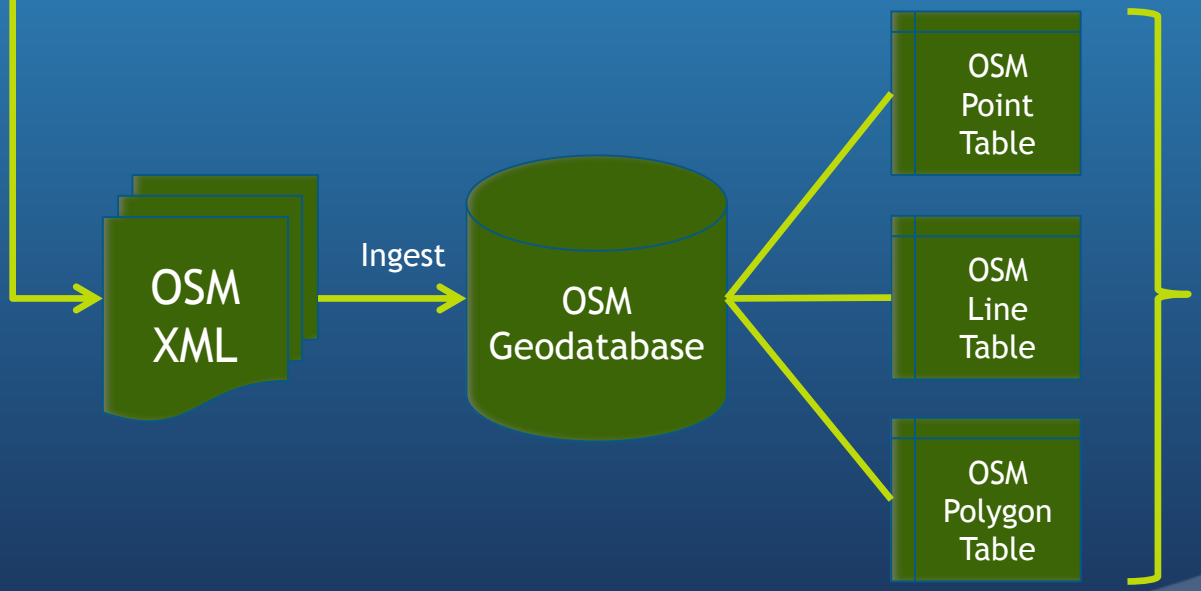
Proposed System Functions

- Information Retrieval System
  - Ingest spatial data – OSM Ingestion Tool
    - Load OSM XML data into a geodatabase
  - Index database – Also part of OSM Ingestion Tool
    - Create index of searchable terms using Lucene Search Engine
  - Linguistically analyze query using NLP tools – Linguistic Analyzer
    - Part of Speech (POS) Tagger
    - Named Entity Recognition (NER)
    - OSM Special Phrases Dictionary
  - Search data – Search Engine
    - Attribute query
    - Spatial query – point/polygon and polygon/polygon relationships
  - Present results – Handed as Search Engine results
    - Expanded Plug-in window shows hits and allows visualization within ArcGIS Desktop map display

Ingest OSM into a Spatial Database

- **OSM Ingestion Tool** – ArcGIS Desktop .Net Plug-in

```
<node id='1667265750' timestamp='2012-03-09T13:30:02Z' uid='621860'
  <tag k='amenity' v='fast_food' />
  <tag k='cuisine' v='burger' />
  <tag k='name' v='McDonald&apos;s' />
</node>
```



Field Name	Data Type
OBJECTID	OBJECT ID
Shape	Geometry
NAME	Text
TYPE	Text
GENERIC	Text
OSMID	Long

Index Spatial Database with Lucene

- Lucene is an Open Source full-text search library written in Java and .Net
- Uses an inverted index providing fast document retrieval
- Higher performance than traditional database SQL search
- Index is stored on file system and can be searched independent of database

ID	…	GENERIC
1	…	name=Starbucks | amenity=cafe | addr:street=Leesburg Pike
2	…	name=Dunkin Donuts | amenity=cafe
3	…	amenity=cafe
4	…	tourism=hotel | name=Homestead
5	…	amenity=fast_food | name=Domino Pizza
6	…	name=Starbucks | amenity=café | addr:street=West Maple Ave.

OSM Database Table

Indexer Pipeline

Token	Documents
amenity=cafe	1, 2, 3, 6
name=Starbucks	1,2
name=Dunkin Donuts	2
tourism=hotel	4
name=Homestead	4
name=Domino Pizza	5
amenity=fast_food	5
addr:street=Leesburg Pike	1
addr:street=West Maple Ave.	6

Inverted Index

Linguistically Analyze Query String

- **Linguistic Analyzer** – ArcGIS Desktop Java plug-in
  - Parse query string (e.g., Starbucks in Vienna)
  - Determine query type: attribute or spatial
    - Spatial preposition IN or ON suggests a spatial query
    - Otherwise, attribute query
  - Determine feature types participating in query
    - Initially limited to points and polygons
  - Identify named entities and parts of speech using Stanford's coreNLP Java library
    - NER Module: Organizations, Locations (e.g., Starbucks, Vienna)
    - POS Module: Prepositions, Nouns (e.g., in, Starbucks, Vienna)
  - Populate Query Object to pass on to the Search Engine

Linguistic Analyzer (cont.)

- Query string “Starbucks in Vienna”
  - POS Tagger: Starbucks/NNP in/IN Vienna/NNP
  - NER: Starbucks [ORGANIZATION] in [OTHER] Vienna [LOCATION]
  - Located spatial preposition IN → Spatial Query
  - Tokenize query string into phrases before and after preposition
    - Left side → Starbucks; Search Lucene index for points and polygons
    - Right Side → Vienna; Search Lucene index for polygons (only)
  - How to construct the search?
    - Named Entities (Organizations and Locations) are likely stored in a name tag as in name=Starbucks and name=Vienna
    - Unmatched entities are checked against Special Phrase Dictionary

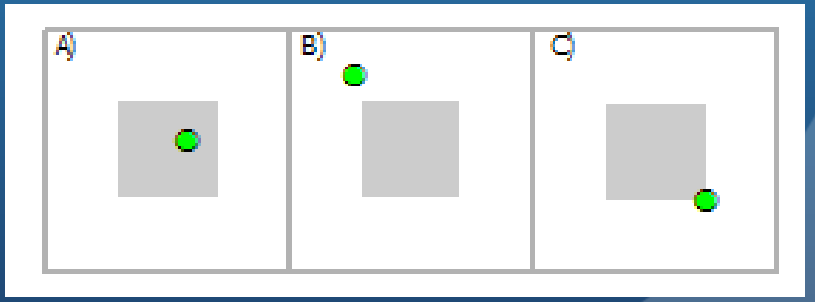
Linguistic Analyzer (cont.)

- Special Phrases Dictionary
  - Built using OSM's Nominatim User Contributed Special Phrases
  - Maps common OSM tag values to fully expanded search strings
    - cafe → amenity=cafe
    - hotels → amenity=hotel
- Query Object populated with parameters for Search Engine
  - Could be multiple objects depending on index search results

PARAMETER	VALUE
QUERY TYPE	SPATIAL
SPATIAL PREPOSITION	IN
SOURCE FEATURE CLASS	POINT
SOURCE SELECTION STRING	Name-Starbucks
TARGET FEATURE CLASS	POLYGON
TARGET SELECTION STRING	Name-Vienna

Search Engine

- Executes Search/Selection of Features based upon:
  - Parameters provided in Query Object
  - For spatial searches, which topological relationship is expressed by the user?
    - Ambiguity in language → What is really meant by **IN**?
    - True for both point/polygon and polygon/polygon relationships
    - For the "Starbucks in Vienna" example, which figure could it be? Does it matter?
    - If B represents a Starbucks on the outskirts of Vienna, does the user want to see it?
- There is a difference in ArcGIS Desktop spatial relationships for the graphic
  - INTERSECT
  - WITHIN
  - COMPLETELY WITHIN
  - HAVE_THEIR_CENTER_IN

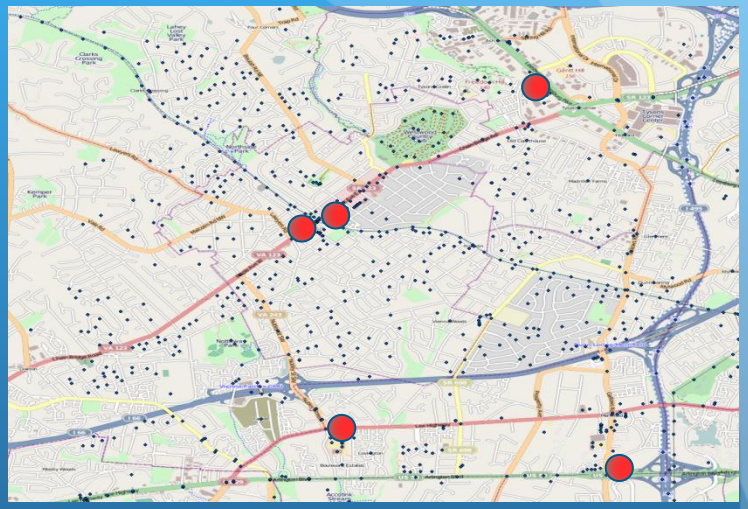


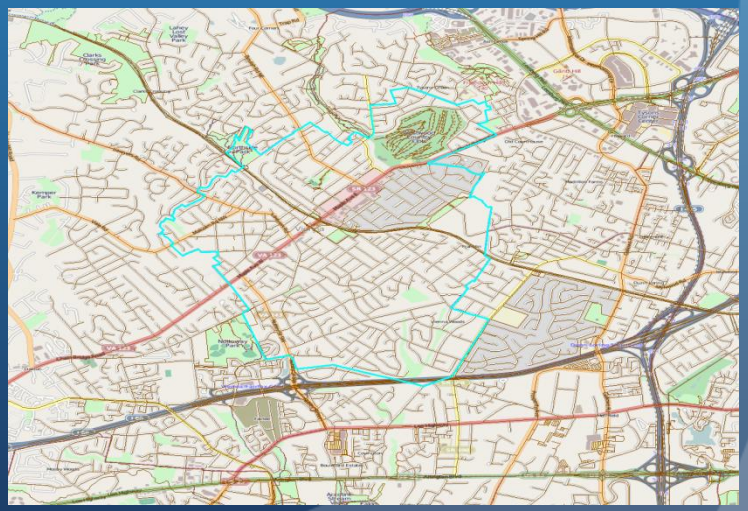
Example – Starbucks in Vienna

- Linguistic Analyzer passes the Query Object to Search Engine

PARAMETER	VALUE
QUERY TYPE	SPATIAL
SPATIAL PREPOSITION	IN
SOURCE FEATURE CLASS	POINT
SOURCE SELECTION STRING	Name-Starbucks
TARGET FEATURE CLASS	POLYGON
TARGET SELECTION STRING	Name-Vienna

- All Starbucks are selected from the POINTS layer
- Polygon Representing Vienna selected from POLYGON layer

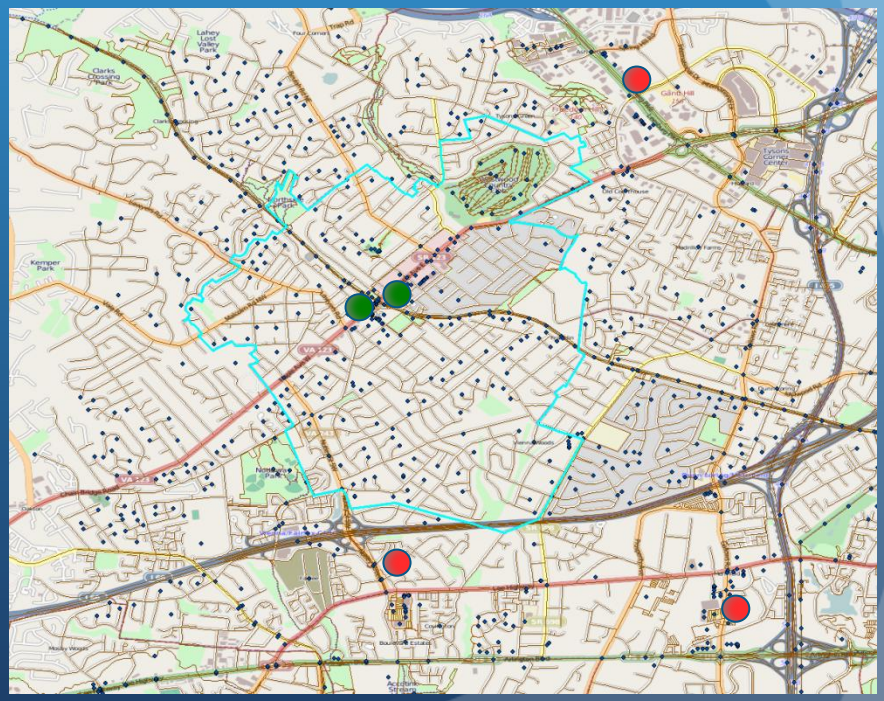




Example – Starbucks in Vienna

- Topological Relationship represented by Query String with clickable results – Only 2 Starbucks are inside the Boundary polygon for Vienna

No	Feature	OSM Key Values
1	Starbucks	amenity:cafe | cuisine:coffee_shop | name:Starbucks
2	Starbucks	amenity:cafe | name:Starbucks



Next Steps and Follow-on Work

- Build the system!
- Determine where I can present my work
- Expand support for additional Spatial Prepositions and more complex query strings
  - Near – need to resolve ambiguity in Near (scale dependency)
  - “Starbucks in Vienna near the airport”
- Expand Query Terms using other NLP Tools and Ontologies
  - Wordnet
- Train NLP Tools on Geographic-term oriented corpora
- Generalize Tool to work with non-OSM data
- Determine how to release code based upon Stanford and OSM Licenses

References

- OpenStreetMap (2012). Copyright and License. Retrieved from http://www.openstreetmap.org/copyright
- Garrod, S., & Coventry, K. (2004). *Saying, Seeing, and Acting: the Psychological Semantics of Spatial Prepositions. Essays in Cognitive Psychology series*. Psychology Press.
- The Stanford Natural Language Processing Group (2012). Retrieved from http://nlp.stanford.edu/software/index.shtml
- Alphabetical list of part-of-speech tags used in the Penn Treebank Project (2012). Retrieved from http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
- OpenStreetMap Wiki (2012). Nominatim. Retrieved from http://wiki.openstreetmap.org/wiki/Nominatim
- ESRI ArcGIS 10.0 Help: Select by location- graphic examples (2012). ESRI, Redlands, California. Retrieved from http://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#//0017000000tp000000
- Kowalski, G. (2011). Information Retrieval Architecture and Algorithms. (G. Kowalski, Ed.) *Information Retrieval*. Springer US. Retrieved from http://www.springerlink.com/content/v0q005/#section=825540&page=1

Questions?