Objects in Python

Methods in Computational Linguistics I
October 15, 2010
Last Time

- List Comprehensions and Dictionaries
Today

- Information Retrieval
- Advertisement delivery
- Objects in Python
Information Retrieval

- Searching for documents, or information within documents.
- Search Engines -- Google, Yahoo, Bing, etc.
- Library search products
- Travel sites -- Orbitz, Expedia, Kayak, etc.
- Job search -- Monster, Careers.com
IR framework

- The user requests information via a **query**
- The system delivers a set of documents, typically ordered by **relevance**.

![Flowchart diagram](chart.png)
Inverted Index

• How can we determine which documents contain the words that we are interested in?
• A - “Cars need gasoline.”
• B - “Gasoline prices rising.”
• C - “Click for prices.”

• cars - {A}
• need - {A}
• gasoline - {A,B}
• prices - {B,C}
• rising - {B}
• click - {C}
• for - {C}
Near misses

- “cars” doesn’t match “car” in an inverted index.
- “running” doesn’t match “run”.
- “running” doesn’t match “marathon”.
Morphological Analysis

• Morphological analysis that converts “cars” to “car +plural”, and “running” to “run +gerund”

• Only store the stem (or lemma) of every word in the index.

• At query time, stem the query.

• CON: This eliminates valuable information from the query.
Query Expansion

- Augment the query with related words, including stems, synonyms, etc.
- similar("running") = {"run", "runs", "ran", "marathon", "race", ...}
- Identifying similar words is an open research question.
Calculating Relevance

• How is relevance calculated?
• A - “Cars need gasoline.”
• B - “Gasoline prices rising.”
• C - “Click for prices.”
• query - “price of gasoline”

• Count the number of hits.
• Count the number of close hits.
• Scale the value of matching a word based on the rarity of the word. Matching “the” is less important than matching “centennial”.
PageRank

• Core of Google’s success in search ~15 years ago.

• Trust a page that people trust. “Crowdsourcing”

• The PageRank of a page is the sum of the PageRank of every page that links to it (divided by the number of links from that page).

\[ PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)} \]
Advertisement Delivery

- Placing a relevant ad on a page can be viewed as an Information Retrieval task.
- Treat the page as the query.
- Treat the set of available ads as documents.
- Relevance needs to involve the price that advertisers will per impression (CPI) or expected payment per click.
- Often, threshold candidate ads based on relevance, then hold an auction to determine the ad to show.
Objects in Python

- Object-Oriented Programming.
- “Objects” are ways of organizing data and the ways that data can be processed.

**Objects** have:

- Variables -- Containing data
- Methods -- Defining how to access and manipulate that data
Objects in NLTK

- We’ve already seen some objects.
- **FreqDist** is an object defined in NLTK.
  - It is a specification of a dictionary, which has some additional functionality.
  - plotting, incrementing the value of elements
- **nltk.Text** is an object
Objects vs. Types

• str and int are special objects.
• They have methods that are defined specifically for them.
• However, these are called **types** because they are built-in to the python language.
• This distinction gets quite blurry.
Example Objects

- Demo time.
- Basic Objects
- Initializing data in an object.
- Objects that can be iterated over.
Inheritance

• Objects can define relationships between objects.

• Through member variables, we can incorporate “has-one” and “has-many” relationships.

• Through **inheritance** we can support “is-a” relationships.
Shape Example

- **Rectangle**
  - area = $h \times w$

- **Triangle**
  - area = $0.5h \times w$

- **Circle**
  - area = $\pi r^2$

- **Square**
  - $h = w$
Next Time

- Regular Expressions in Python