Hunter Gatherer

1-Click Search Summarizer – Université de Montréal

Who?

- GRIUM
  - Groupe de Recherce de Information à UdeM
- PI
  - Prof. Jian-Yun Nie
- Team
  - Dr. Pablo Duboue (presenting)
  - Dr. Jing He

What?

- 1-Click Search
    - Input: Query and 200 ranked Web pages
    - Output: a 1,000 characters summary
        - Summary should contain the information the pages relevant to the query.
- A research challenge part of NTICR
- Queries belong to 8 types (celebrities, how to, location, etc)
    - But the type is not explicit

http://research.microsoft.com/en-us/people/tesakai/1click2.aspx

Examples

- Query: Whitney Houston Death
- Relevant information:
    - On February 11, 2012, Houston was found dead in suite 434 at the Beverly Hilton Hotel, submerged in the bathtub.
        - V001001 February 11, 2012
        - V001002 Beverly Hilton Hotel
        - V001003 suite 434
        - V001004 submerged in the bathtub
    - the cause of Houston's death was drowning and the "effects of atherosclerotic heart disease and cocaine use".
        - V002001 drowning
        - V002002 atherosclerotic heart disease
        - V002003 cocaine

Actual Queries

1C2-E-0001	michael jackson death
1C2-E-0002	marvin gaye influence
1C2-E-0004	dr der
1C2-E-0005	keith sweat
1C2-E-0006	glen campbell
1C2-E-0007	whitney houston movies
1C2-E-0008	Lil wayne
1C2-E-0009	john denver
1C2-E-0010	rodney atkins
1C2-E-0017	joe arroyo
1C2-E-0022	selena gomez
1C2-E-0025	marlon brando acting style
1C2-E-0026	jennifer gardner alias
1C2-E-0038	james cameroon biggest movies
1C2-E-0042	robert kennedy cuba
1C2-E-0045	mayor bloomberg
1C2-E-0046	19th president us
1C2-E-0047	tom corbett
1C2-E-0048	nancy pelosi
1C2-E-0049	ron paul tea party
1C2-E-0050	mitt romney governor ma

Hunter Gatherer Approach

- Apply the DeepQA architecture to 1-Click task
  - Do not explicitly type the query
- Hunt nuggets, gather evidence

1. Hunt text nuggets on relevant passages
2. Gather evidence passages that contain nuggets and query terms
3. Score nuggets based on evidence
4. Final output are sentences containing highly scored nuggets

https://github.com/jinghe/hunter-gatherer

Hunter Gatherer Approach

1.Hunt text nuggets on relevant passages

- On [February] [11], [2012], [Houston] was [found dead] in [suite 434] at the [Beverly Hilton Hotel], submerged in the [bathtub]

1.Gather evidence passages that contain nuggets and query terms

- Query: Whitney Houston “suite 434“

1.Score nuggets based on evidence

2.Final output are sentence containing highly scored nuggets

Results

- Mixed results
  - Spam and repeated passages were our doom
- Query: Hilary Clinton first lady
- of America North America Turkey First Ladies Visits Airports Jeeps vehicles speeches Addresses Sermons Applause Clapping Motor Cars Cars Motorcars Women Hillary Clinton Bill Clinton Wife Hilary Clinton Hillary Rodham Clinton Istanbul Cirgan Palace Hotel Istanbul. Fraser on September 7, 2011 | Leave Comments | Related : Hillary Clinton, Michelle Obama, Prince Harry, Tabloid Wednesday. Clinton was elected to the United States Senate in 2000, becoming the first First Lady elected to public office and the first woman elected statewide in New York. Hillary Diane Rodham Clinton (born October 26, 1947) is the 67th United States Secretary of State, serving in the administration of President Barack Obama. She is married to Bill Clinton, the 42nd President of the United States, and was the First Lady of the United States from 1993 to 2001. Globe claims: Hillary Clinton suffered an alarming secret breakdown after a bitter clash with First Lady Michelle Obama, and now Bill

Why Python

- Hunter-Gatherer uses Python true to its duct tape origins
  - Two people working closely together
  - Very tight deadline
  - Integrating large number of existing tools and libraries
    - INDRI, NLTK, CCLParser, Glpk, Mallet, Boilerplate
  - Very exploratory coding
  - Code is the only documentation

Case Study: GLPK

- A state of the art summarizationtechnique involves using Integer Linear Programming and expressing the selection of sentences as an optimization problem
    - There are N nuggets and M sentences
    - Some sentences contain some nuggets
    - Each nugget has an score
    - We want to select sentences up to a certain length so to maximize the scores of the contained nuggets

GPLK: GNU Linear Programming Kit

- A DSL for linear programming
  - param NS; param NC; param K;
  - param M{1..NS, 1..NC}, binary; param L{1..NS}, integer; param W{1..NC} ;
  - var s{1..NS}, binary; var e{1..NC}, binary;
  - maximize z: sum { i in 1..NC } e[i]*W[i];
  - subject to l:
    - sum { i in 1..NS } L[i]*s[i] <= K;
  - subject to m {j in 1..NC}:
    - sum { i in 1..NS } M[i,j]*s[i] >= e[j];

import gplk

```
constraints = glpk.glpk(f.name)
constraints.update()
constraints.solve()
evidence = list()
for sent_idx in xrange(len(sentences)):
    if constraints.s[sent_idx+1].value() == 1.0:
      evidence.append(sentences[sent_idx][2])
```

Concluding rants

- NLTK “rant“ – parser.py:100 def mix_brackets ...
- Personal “rant“

https://github.com/jinghe/hunter-gatherer