U.S. National Park Service: Exploring Visitor Data

Andrea Everett • 11.11.2016

Data

NPS visitor questionnaires (2016)

- 320 unique park units

- 40, 380 free-form visitor responses (~20-250 per unit)

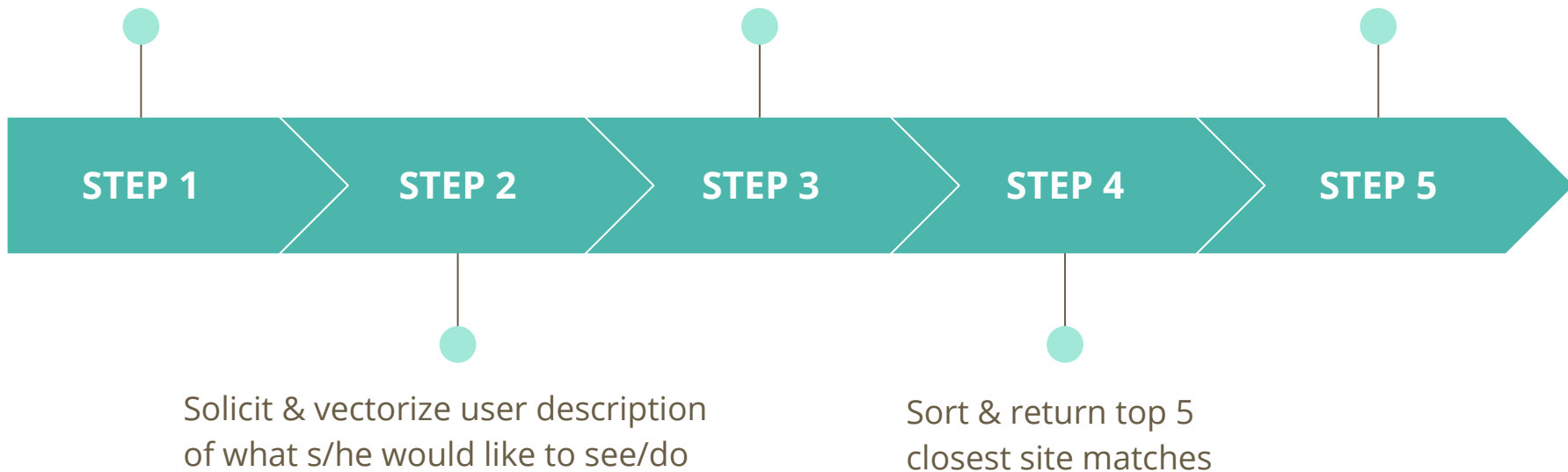- How did the visitor understand the significance of the park unit?

Questions

- How do people feel about the park sites they visit?

- How do they understand the unique significance of these places?

- Can we use their experiences to make recommendations to other potential visitors?

1. Recommendation system (content-based)

Use word2vec & Google word vectors to create an average vector for each NPS site

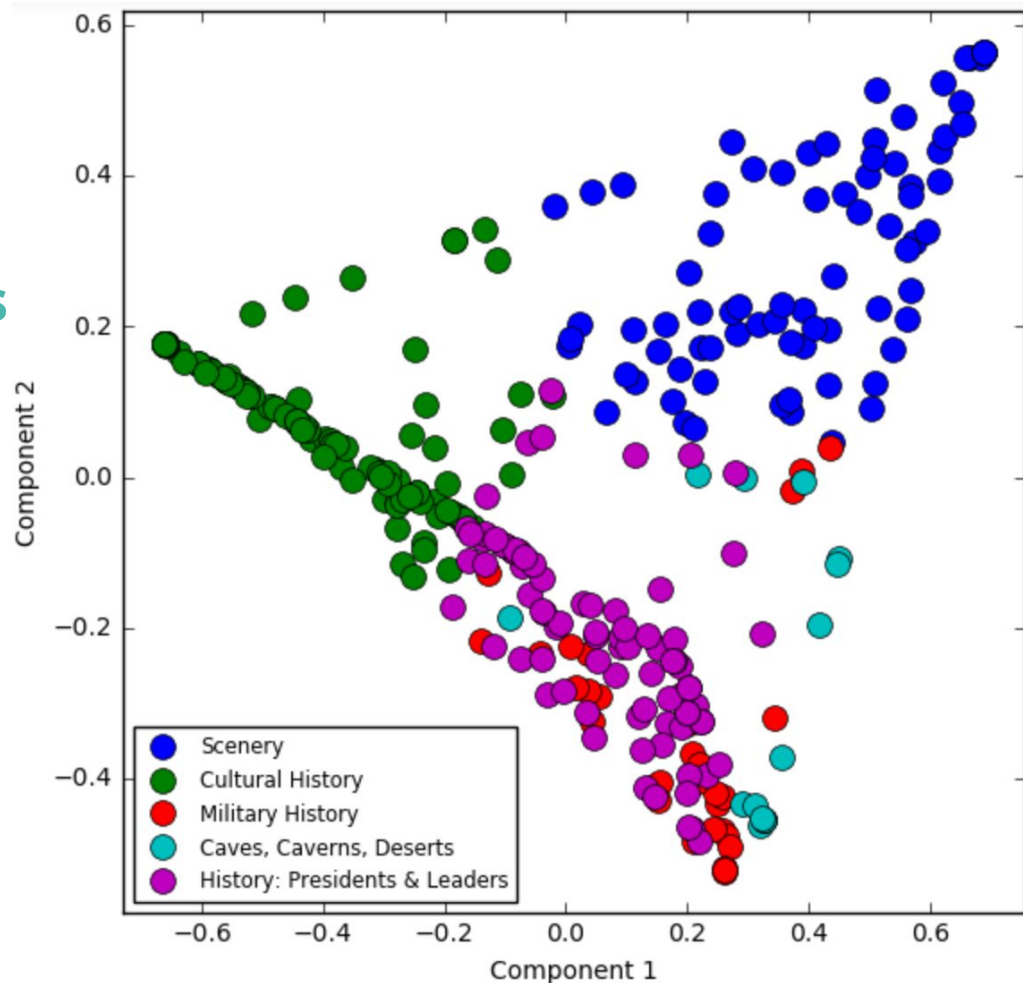Compare cosine similarity of user vector with all 320 site vectors

Flask App Demonstration!

STEP 1   **STEP 2**   **STEP 3**   **STEP 4**   **STEP 5**

Solicit & vectorize user description of what s/he would like to see/do

Sort & return top 5 closest site matches

2. Topic Modeling

Hierarchical Dirichlet Process

- 150 topics

- Visualized with pyLDAvis
  (Sievert & Shirley 2014)

Kmeans clustering

- 5 clusters

- PCA: reduced to 2 dimensions

3. TFIDF

What unique terms did visitors use to describe each site?

- TfidfVectorizer

- All comments, merged for each site

- Represented with word clouds

Great Basin National Park

Ellis Island

Hawaii Volcanoes National Park

4. Sentiment Analysis

Which park sites do visitors feel most positively about?

- TextBlob

- All comments, averaged by site

- Overall sentiments not very strong

ParkName	polarity
Great Sand Dunes NP	0.329459
Big South Fork NRRA	0.329201
Abraham Lincoln Birthplace NHS	0.322277
Carlsbad Caverns NP	0.317445

Thank you

NATIONAL PARK SERVICE

Questions?