Building a Recommender System

A content-based recommendation system for REI's hiking project

Outline

1. Overview and Inspiration for Product

2. Building the Recommendation System - From Wrangling to Evaluation

3. How well did these perform? An Evaluation

4. Who benefits and how could this improve?

So many hikes, so little time

Colorado	4,983 Trails
Divide	41 Trails
Ft Collins	76 Trails
Guanella Pass	8 Trails
Roosevelt National Forest	404 Trails
Steamboat Springs	68 Trails
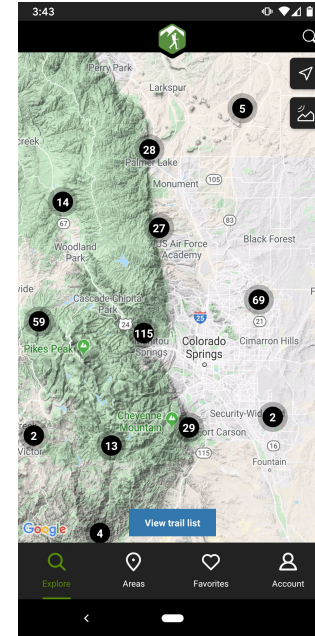
Overview And Inspiration ◯ Building the Recommender System ◯ How well does this perform? ◯ Who Benefits and how could this improve?

How would the recommender system work?
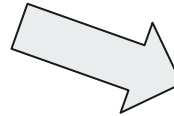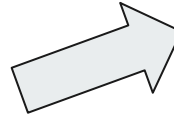
Filter Results - API Request

- User's Location

- Drive Distance

- Max Hike Length

Clean and Organize

- Text Cleaning

- Handle Missing Values

- Create Features

Keyword Recommender
10 best hikes where summary matches user defined keywords

Similar Hike Recommender
10 best hikes that match a target or pre-defined hike

Overview And Inspiration

Building the Recommender System

How well does this perform?

Who Benefits and how could this improve?

Wrangling Data

Wrangle data for exploration purposes. REI's Hiking Project API needs a few parameters:

1) Latitude
2) Longitude
3) Maximum Distance
4) Number of Records to Return

Overview
And
Inspiration

**Building the
Recommender
System**

How well does
this perform?

Who Benefits
and how could
this improve?

What are the details?

- 12,764 unique hikes

- Some features are inverses (high/low and ascent/descent)

- Median rating is a 4 out of 5

- Average hike length is about 5 miles -- largest goes to 195 miles

Overview
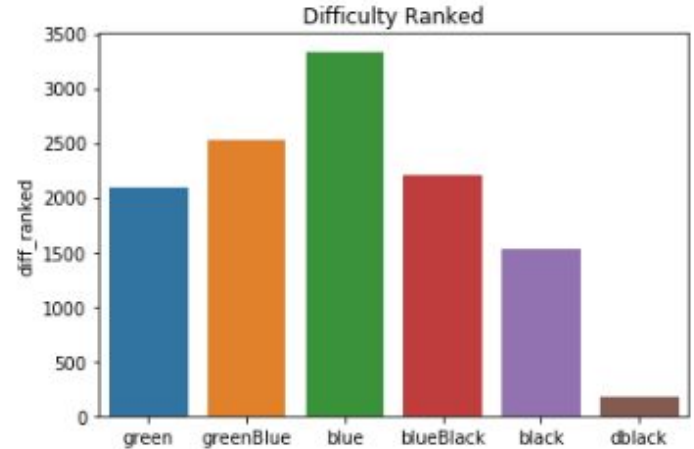And
Inspiration

**Building the
Recommender
System**

How well does
this perform?

Who Benefits
and how could
this improve?

Details Continued..

Difficulty Ranked

- **Some features entirely missing (condition details)**

- **Difficulty evenly spread**

- **Summary feature uses regular English (no emojis, internet slang, etc.)**

Overview
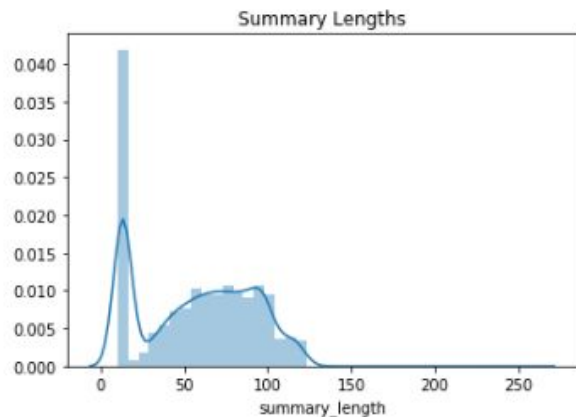And
Inspiration

Building the Recommender System

How well does this perform?

Who Benefits and how could this improve?

Odd summary lengths that need handling

Summary Lengths

needs summary	2698
needs adoption	347
needs adoption!	26
coming soon	11
adopt me, please!	9
adopt this trail!	8
to be written	7
needs adoption.	6
needs summary.	6
to be written.	2

Overview
And
Inspiration

**Building the
Recommender
System**

How well does
this perform?

Who Benefits
and how could
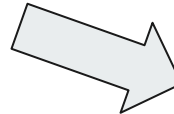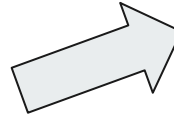this improve?

How would the recommender system work?

Filter Results - API Request

- User's Location

- Drive Distance

- Max Hike Length

Clean and Organize

- Text Cleaning

- Handle Missing Values

- Create Features

Keyword Recommender
10 best hikes where summary matches user defined keywords

Similar Hike Recommender
10 best hikes that match a target or pre-defined hike

Overview And Inspiration

Building the Recommender System

How well does this perform?

Who Benefits and how could this improve?

Creating Our Filtered Dataset

1) We need the latitude/longitude of the user

   a) Found a dataset of ~2,500 most populated US cities
   b) Fuzzy string matching that returned the best matches' lat/long

2) User input for maximum willingness to drive

3) Return max-results (500) then filter for maximum trail length

Overview
And
Inspiration

**Building the
Recommender
System**

How well does
this perform?

Who Benefits
and how could
this improve?

Clean Our Filtered Dataset

1) Make our summary all lowercase

2) Remove our summaries that have no actual values

3) Remove all rows with null values

4) Create feature needed for our bag-of-words

```
south creek washington : 347 records
ferry pass florida : 77 records
ozona texas : 71 records
lewisburg tennessee : 394 records
elim pennsylvania : 398 records
montgomery alabama : 291 records
duluth georgia : 390 records
chestertown maryland : 414 records
olympian village missouri : 268 records
fontana kansas : 286 records
colome south dakota : 1 records
```

Overview
And
Inspiration

**Building the
Recommender
System**

How well does
this perform?

Who Benefits
and how could
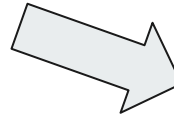this improve?

How would the recommender system work?

Filter Results - API Request

- User's Location

- Drive Distance

- Max Hike Length

Clean and Organize

- Text Cleaning

- Handle Missing Values

- Create Features

Keyword Recommender
10 best hikes where summary matches user defined keywords

Similar Hike Recommender
10 best hikes that match a target or pre-defined hike

Overview And Inspiration

Building the Recommender System

How well does this perform?

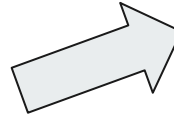Who Benefits and how could this improve?

Similar Hike Recommender

1) Uses bag-of-words count vectorizer of the summary feature

2) Also uses other features (stars, length, ascent, descent, high, low)

3) Computes cosine-similarity between filtered and the 'target' hike
   a) Note: Model also performed with euclidean distances -- not as effective

4) Returns the top ten best matches

Overview
And
Inspiration

**Building the
Recommender
System**

How well does
this perform?

Who Benefits
and how could
this improve?

Keyword Matching Recommender

1) Uses filtered dataset and perform NLP on our summary (no stop-words, no punctuation, etc.)

2) Has users input word(s) to be used as keywords

3) Uses Word2Vec function to compute vector similarity between NLP summary and keywords

4) Returns best 10 matches

Overview
And
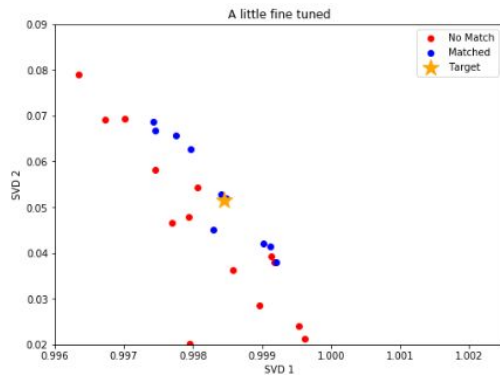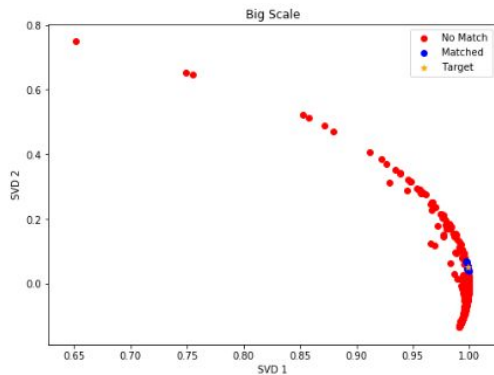Inspiration

**Building the
Recommender
System**

How well does
this perform?

Who Benefits
and how could
this improve?

How well does our similar hikes model perform?

Overview And Inspiration

Building the Recommender System

How well does this perform?

Who Benefits and how could this improve?

Denver, Colorado -- Keywords: 'Easy'..'Water'

this is a great place to camp on a hot summer day. come for the views, but stay for the cool water and the upper lake.

a fairly easy hike providing awesome views from a high alpine lake.

quickly earn some amazing mountain views and slowly gain an alpine lake on this easy-to-access route

starting from a less-crowded lot, this offers rougher terrain and easy climbing access. great views of downtown golden.

a solid hike with a good mix of open space and tree cover, single and doubletrack trail, and a couple of scenic views.

a good hike with easy front range access. some steep climbs, rocks and nice views.

fairly easy trail to a beautiful place. continue up to the ridgeline for even better views.

a relatively easy trail with beautiful views of lakes and peaks.

a fun trail to let your dog off the leash with many water crossings to romp through.

a two-faced (easy, then very hard) trail that leads to a stunning waterfall and serene alpine lake.

| Overview And Inspiration | ◯ | Building the Recommender System | ◯ | **How well does this perform?** | ◯ | Who Benefits and how could this improve? |

How well does our keyword(s) model perform?
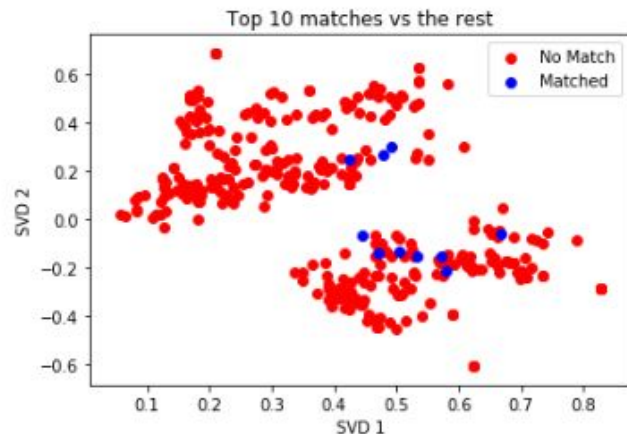
Top 10 matches vs the rest

Overview
And
Inspiration

Building the
Recommender
System

**How well does
this perform?**

Who Benefits
and how could
this improve?

Who can benefit from these models?

1) The user of app (both web and mobile)

2) REI could use this for marketing reasons

3) Increase valid summaries by recommending hikes that need summaries and are similar (can no longer use bag-of-words)

Overview And Inspiration

Building the Recommender System

How well does this perform?

Who Benefits and how could this improve?

How can our models be improved?

1) Create a collaborative-filtering recommender (need user data)

2) Get more summaries and train unique Word2Vec Model for keyword improvement

3) Increase number of steps? Potentially increase computational resources

Overview
And
Inspiration

Building the
Recommender
System

How well does
this perform?

**Who Benefits
and how could
this improve?**

Similar Hikes - Clustering First Method

- **Same preliminary steps - filtered subset**

- **Performs K-Means Clustering to create another subset**

- **Performs our similar hikes on this set**
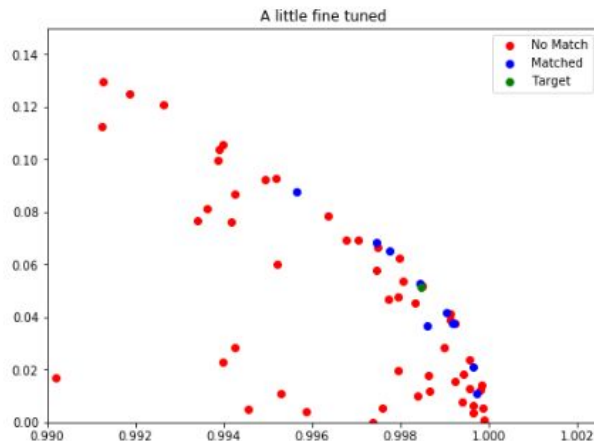
A little fine tuned

Overview And Inspiration

Building the Recommender System

How well does this perform?

Who Benefits and how could this improve?

Appendix and Reference Points

1) Map of area to cover for EDA

2) Duplicate Example

3) Summary Examples

Which areas to cover?

Pacific Northwest

Colorado
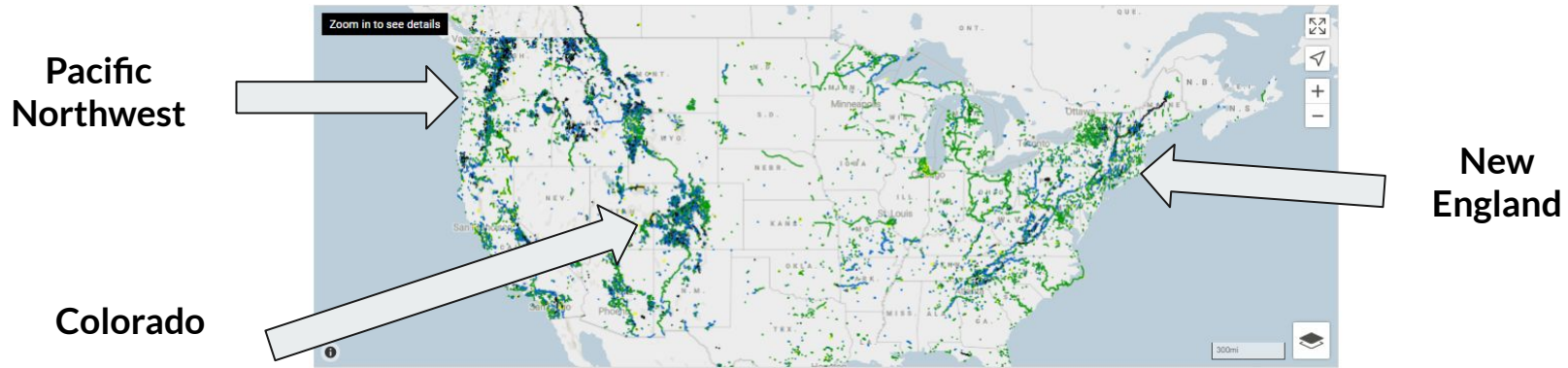
New England

Zoom in to see details

Overview And Inspiration

Building the Recommender System

How well does this perform?

Who Benefits and how could this improve?

Removing Duplicates

	id	name	summary	difficulty	stars	star_votes	location	url	length	ascent	descent	high	low	longitude	latitud
0	7003796	Section 16/Palmer Loop	A favorite loop in the Springs area.	blue	4.5	72	Manitou Springs, Colorado	https://www.hikingproject.com/data/	5.9	1287	-1288	7845	6633	-104.8903	38.822
1	7003796	Section 16/Palmer Loop	A favorite loop in the Springs area.	blue	4.5	72	Manitou Springs, Colorado	https://www.hikingproject.com/data/	5.9	1287	-1288	7845	6633	-104.8903	38.822
2	7003796	Section 16/Palmer Loop	A favorite loop in the Springs area.	blue	4.5	72	Manitou Springs, Colorado	https://www.hikingproject.com/data/	5.9	1287	-1288	7845	6633	-104.8903	38.822
3	7003796	Section 16/Palmer Loop	A favorite loop in the Springs area.	blue	4.5	72	Manitou Springs, Colorado	https://www.hikingproject.com/data/	5.9	1287	-1288	7845	6633	-104.8903	38.822
4	7003796	Section 16/Palmer Loop	A favorite loop in the Springs area.	blue	4.5	72	Manitou Springs, Colorado	https://www.hikingproject.com/data/	5.9	1287	-1288	7845	6633	-104.8903	38.822

Overview And Inspiration

Building the Recommender System

How well does this perform?

Who Benefits and how could this improve?

What about the summary feature?

- **Straightforward English**

- **No Emojis/internet language**

a favorite loop in the springs area.
a fun, rolling trail with a sense of seclusion, great geology, and historical sites.
a longer loop in this spectacular area allows hikers to take in the highlights without the crowds.
a truly beautiful hike following ice cave creek and levi creek, then following a ridge to the harrison plane crash.
easily accessible from colorado spings, with views of the city and 360° mountain views from the top.
an excellent, but short, loop through the heart of garden of the gods.
a hike with creeks, bridges, scree slopes, forested paths, summit to mount rosa, then finishing with st. mary falls.
a solid ascent along the ridge northwest of the upper reservoir and descending the old winding staircase-beautiful views
widely recognized as one of the best urban trails in the country.
this 750 acre park features geological formations including spires, chasms, and gullies.

Overview And Inspiration

Building the Recommender System

How well does this perform?

Who Benefits and how could this improve?