Job Similarity Recommendation

Craig Johnson

Overview

- Project description
- Introduction
- Dataset creation
- Model Building
- Results
- Exploratory model comparison
- Discussion

Project description

Project Description

- Using text, predict what jobs are most and least similar in the O*NET database
- Verify if recommendations have face validity
- Verify that results align with expected Holland codes
- Verify break the world of work into six categories

Introduction

Occupational Information Network (O*NET)

- Developed in 1998 to replace the Dictionary of Occupational Titles
- Used by the government and the general workforce as the major reference on jobs
- Each job in the US workforce described in a standardized way using the O*NET
- Content Model
    - Worker Characteristics
    - Worker Requirements
    - Experience Requirements
    - Occupational Requirements
    - Workforce Characteristics
    - Occupation-Specific Information

O*NET continued….

- Each job is labeled with a unique O*NET SOC code
- The O*NET SOC code is hierarchical in nature with major, minor, broad, and detailed occupations.
- <u>Take-away</u>: Similar jobs should have similar O*NET Codes.

Holland Codes

- Created by Dr. John Holland in the 1950’s
- Breaks the world of work into six categories
    - Realistic
    - Investigative
    - Artistic
    - Social
    - Enterprising
    - Conventional
- Used to help match people’s interests to jobs and education
- Take-away: Jobs that are similar should have similar Holland codes

Dataset Creation

Dataset Creation

1. Review O*NET Content model to determine relevant textual components
2. Review O*NET online to identify six sections and verify sections
3. Review O*NET database to identify appropriate variables
4. Review O*NET data dictionary to identify appropriate variable values
5. Query identified data sections and place into pandas dataframes
6. Merge data section together into a single dataset
7. Concatenate the “stacked” text for each job into a single long string
8. “Preprocess” text putting all text into lowercase, tokenizing, and remove english stop words

Example of the final dataset…

	onetsoc_code	First Interest High-Point	Second Interest High-Point	Third Interest High-Point	riasec	title	text
0	11-1011.00	Enterprising	Conventional	None	EC	Chief Executives	Direct or coordinate an organizations financia...
1	11-1011.03	Enterprising	Conventional	Investigative	ECI	Chief Sustainability Officers	Identify educational training or other develop...
2	11-1021.00	Enterprising	Conventional	Social	ECS	General and Operations Managers	Direct and coordinate activities of businesses...
3	11-1031.00	Enterprising	Social	None	ES	Legislators	Analyze and understand the local and national ...
4	11-2011.00	Enterprising	Artistic	Conventional	EAC	Advertising and Promotions Managers	Prepare budgets and submit estimates for progr...

Model Building

Word2Vec

- Form of textual analysis
- Uses a shallow neural network to generate word vectors
- Word vectors create a vector of weights which can be used to make predictions and determine similarity between words
- Take-away: Jobs that are similar should have vectors which are similar. In mathematical terms, they will have a cosine similarity near one.

A vector example...

Words that are similar will have a smaller cosine between them. Values near 1 indicate high similarity while 0 indicate no similarity.

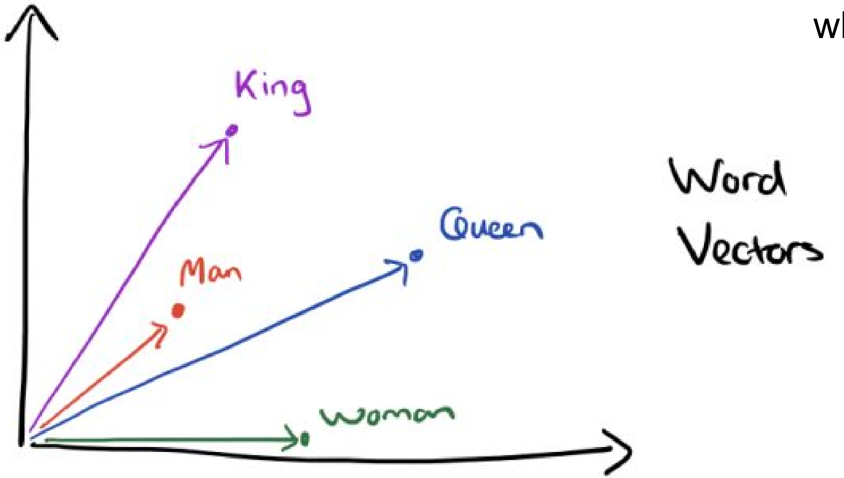


Image from https://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/

Model building....

1. GoogleNews-vectors-negative300
    a. Downloaded at https://code.google.com/archive/p/word2vec/
2. O*NET
    a. Model creation software: Gensim 3.0 (word2vec model)
    b. Settings
        i. Size = 300
        ii. Window width = 5
        iii. Iterations = 15
3. Take-away: Two trained models
    a. Model 1: Google trained
    b. Model 2: Model trained on O*NET

Average word embeddings

- ***Problem***: Word2Vec calculates word embeddings for a single word, how do we handle the long strings with all the job description data from O*NET?
- ***Solution***: Average word embeddings
  - Calculate the word embeddings for each word in the description and then average each element together. Each job then has a single 300 dimension set of values
- Steps
  - Calculate average word embeddings <u>for each job</u> using the Google word vectors, save the dataset.
  - Calculate average word embeddings <u>for each job</u> using the O*NET word vectors, save the dataset.

Example of the final dataset output

	title	d0	d1	d2	d3	d4	d5	d6	d7	d8	d9
379	Music Directors	0.026201	0.024999	-0.017454	-0.012312	-0.047884	-0.003302	0.020082	0.011222	-0.035705	0.032271
378	Choreographers	0.027697	0.012169	-0.002602	-0.010429	-0.045809	-0.000181	0.009114	0.016606	-0.034702	0.033165
384	Public Address System and Other Announcers	0.012593	0.017254	-0.017846	-0.014706	-0.033975	0.007945	0.010145	0.011816	-0.027836	0.025319
368	Actors	0.041632	0.021967	0.001624	-0.013410	-0.054547	0.024520	-0.013511	0.023597	-0.052994	0.040552
372	Talent Directors	0.008218	0.023569	-0.009766	-0.013399	-0.055430	0.006866	0.004151	0.021393	-0.053137	0.012393

Results

Results: Chefs and Head Cooks

ONET:35-1011.00
Holland Code:ERA
****
Google data
****
The most similar jobs are...

- Cooks, Institution and Cafeteria; cosine:0.99; O*NET:35-2012.00; Holland code:RC
- Cooks, Short Order; cosine:0.99; O*NET:35-2015.00; Holland code:RC
- Combined Food Preparation and Serving Workers, Including Fast Food; cosine:0.99; O*NET:35-3021.00; Holland code:CRE
- Cooks, Restaurant; cosine:0.99; O*NET:35-2014.00; Holland code:RE
- Cooks, Private Household; cosine:0.99; O*NET:35-2013.00; Holland code:ARC
- Cooks, Fast Food; cosine:0.99; O*NET:35-2011.00; Holland code:RC
- Baristas; cosine:0.98; O*NET:35-3022.01; Holland code:ECR
- Counter Attendants, Cafeteria, Food Concession, and Coffee Shop; cosine:0.98; O*NET:35-3022.00; Holland code:RSE
- Food Servers, Nonrestaurant; cosine:0.98; O*NET:35-3041.00; Holland code:SRE
- Food Preparation Workers; cosine:0.98; O*NET:35-2021.00; Holland code:RC

The least similar jobs are...

- Software Developers, Applications; cosine:0.84; O*NET:15-1132.00; Holland code:IRC
- Investment Underwriters; cosine:0.84; O*NET:13-2099.03; Holland code:CE
- Green Marketers; cosine:0.83; O*NET:11-2011.01; Holland code:EAI
- Fuel Cell Technicians; cosine:0.82; O*NET:17-3029.10; Holland code:RCI
- Data Warehousing Specialists; cosine:0.75; O*NET:15-1199.07; Holland code:IC

Results: Chefs and Head Cooks

ONET:35-1011.00
Holland Code:ERA
****
O*NET data
****
The most similar jobs are...

Cooks, Institution and Cafeteria; cosine:0.98; O*NET:35-2012.00; Holland code:RC
Cooks, Private Household; cosine:0.96; O*NET:35-2013.00; Holland code:ARC
Cooks, Restaurant; cosine:0.96; O*NET:35-2014.00; Holland code:RE
Baristas; cosine:0.95; O*NET:35-3022.01; Holland code:ECR
Food Service Managers; cosine:0.95; O*NET:11-9051.00; Holland code:ECR
First-Line Supervisors of Aquacultural Workers; cosine:0.95; O*NET:45-1011.06; Holland code:ERC
Cooks, Fast Food; cosine:0.95; O*NET:35-2011.00; Holland code:RC
Dietetic Technicians; cosine:0.95; O*NET:29-2051.00; Holland code:SIR
First-Line Supervisors of Housekeeping and Janitorial Workers; cosine:0.95; O*NET:37-1011.00; Holland code:ECR
First-Line Supervisors of Landscaping, Lawn Service, and Groundskeeping Workers; cosine:0.95; O*NET:37-1012.00; Holland code:ERC

The least similar jobs are...

Methane/Landfill Gas Collection System Operators; cosine:0.44; O*NET:11-3051.05; Holland code:CER
Green Marketers; cosine:0.43; O*NET:11-2011.01; Holland code:EAI
Methane/Landfill Gas Generation System Technicians; cosine:0.34; O*NET:51-8099.02; Holland code:RCI
Data Warehousing Specialists; cosine:0.31; O*NET:15-1199.07; Holland code:IC
Fuel Cell Technicians; cosine:0.20; O*NET:17-3029.10; Holland code:RCI

Results: Construction Carpenters

ONET:47-2031.01
Holland Code:RCI
****
Google data
****
The most similar jobs are...

- Rough Carpenters; cosine:1.00; O*NET:47-2031.02; Holland code:RCI
- Helpers--Carpenters; cosine:0.99; O*NET:47-3012.00; Holland code:RC
- Brickmasons and Blockmasons; cosine:0.99; O*NET:47-2021.00; Holland code:RCI
- Cabinetmakers and Bench Carpenters; cosine:0.99; O*NET:51-7011.00; Holland code:RC
- Structural Metal Fabricators and Fitters; cosine:0.99; O*NET:51-2041.00; Holland code:RC
- Helpers--Brickmasons, Blockmasons, Stonemasons, and Tile and Marble Setters; cosine:0.99; O*NET:47-3011.00; Holland code:R
- Mechanical Door Repairers; cosine:0.98; O*NET:49-9011.00; Holland code:R
- Drywall and Ceiling Tile Installers; cosine:0.98; O*NET:47-2081.00; Holland code:RC
- Sawing Machine Setters, Operators, and Tenders, Wood; cosine:0.98; O*NET:51-7041.00; Holland code:RCI
- Model Makers, Wood; cosine:0.98; O*NET:51-7031.00; Holland code:RAC

The least similar jobs are...

- Fuel Cell Technicians; cosine:0.83; O*NET:17-3029.10; Holland code:RCI
- Methane/Landfill Gas Collection System Operators; cosine:0.83; O*NET:11-3051.05; Holland code:CER
- Investment Underwriters; cosine:0.80; O*NET:13-2099.03; Holland code:CE
- Green Marketers; cosine:0.79; O*NET:11-2011.01; Holland code:EAI
- Data Warehousing Specialists; cosine:0.75; O*NET:15-1199.07; Holland code:IC

Results: Construction Carpenters

ONET:47-2031.01
Holland Code:RCI
****
O*NET data
****
The most similar jobs are...

Rough Carpenters; cosine:0.98; O*NET:47-2031.02; Holland code:RCI
Brickmasons and Blockmasons; cosine:0.97; O*NET:47-2021.00; Holland code:RCI
Cabinetmakers and Bench Carpenters; cosine:0.96; O*NET:51-7011.00; Holland code:RC
Helpers--Roofers; cosine:0.96; O*NET:47-3016.00; Holland code:RC
Sheet Metal Workers; cosine:0.96; O*NET:47-2211.00; Holland code:R
Roofers; cosine:0.95; O*NET:47-2181.00; Holland code:RC
Explosives Workers, Ordnance Handling Experts, and Blasters; cosine:0.95; O*NET:47-5031.00; Holland code:RIC
Electromechanical Equipment Assemblers; cosine:0.95; O*NET:51-2023.00; Holland code:RCI
Painters, Construction and Maintenance; cosine:0.95; O*NET:47-2141.00; Holland code:RC
Drywall and Ceiling Tile Installers; cosine:0.95; O*NET:47-2081.00; Holland code:RC

The least similar jobs are...

Special Education Teachers, Preschool; cosine:0.23; O*NET:25-2051.00; Holland code:SA
Data Warehousing Specialists; cosine:0.22; O*NET:15-1199.07; Holland code:IC
Green Marketers; cosine:0.20; O*NET:11-2011.01; Holland code:EAI
Legislators; cosine:0.19; O*NET:11-1031.00; Holland code:ES
Investment Underwriters; cosine:0.16; O*NET:13-2099.03; Holland code:CE

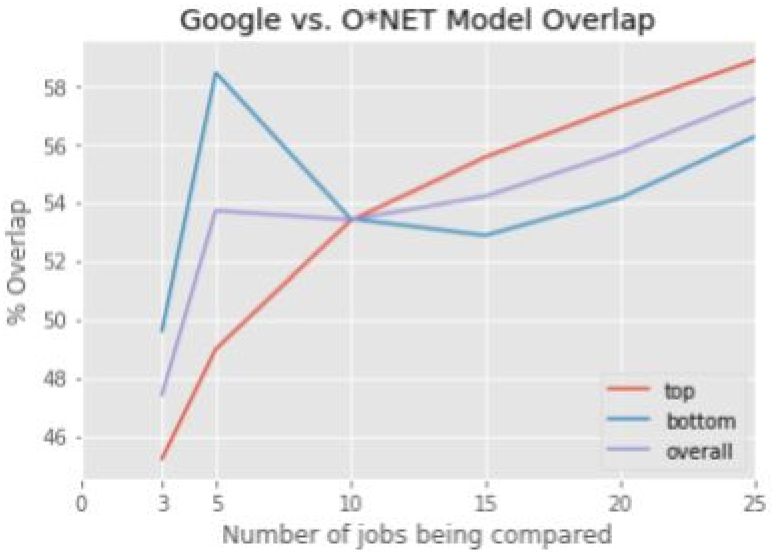
General Results take-away

- Results seemed to have high face validity
- O*NET SOC codes seemed similar to the target job
- Holland codes were generally similar to those of the target job
- Google and O*NET models produced similar results
- Take-aways:
    - Word2vec results make sense
    - Word2vec results align pretty well with the theory outlined by Holland codes
    - Word2vec results align with the O*NET SOC code hierarchy

Exploratory model comparison

Model comparison

- *Question*: How much overlap was there between the recommended jobs using the two models?
-



Discussion

Discussion

- Word2vec provided results that had high face validity and were aligned with O*NET SOC codes as and Holland codes
- What does it mean when they don't align?
  - Bad model? Incomplete or insufficient textual data? Old text? Problems with the specificity of the O*NET model? Bad expert judgment?
- Future directions
  - Explore the incongruencies in more depth
  - Try and create a system to classify jobs automatically