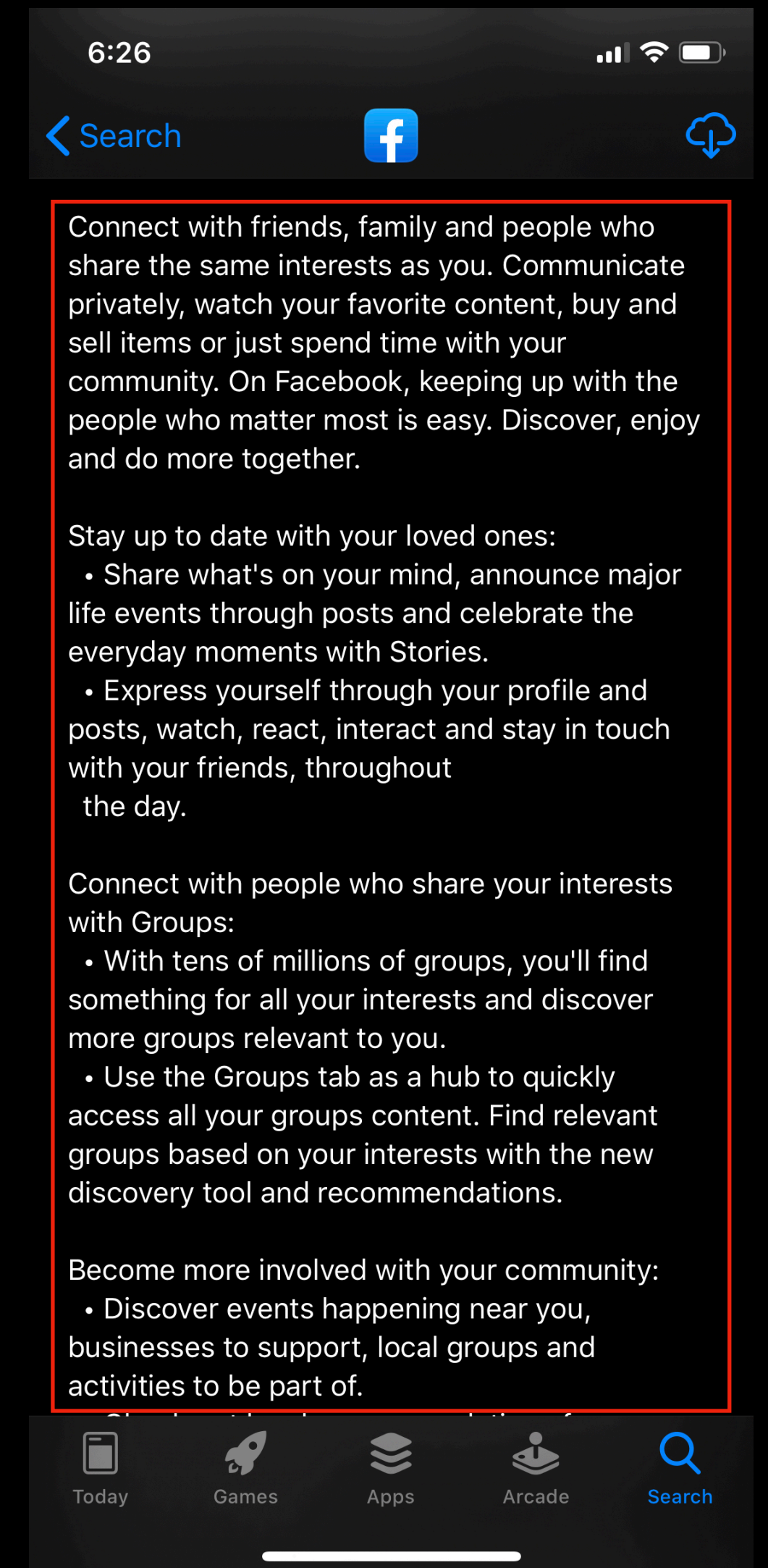# Finding similar apps

COMPSCIX433.6-009 Introduction to Machine Learning Using Python Project

Souvik Roy 06/15/20

Goal
Finding similar apps

- Learn from App Description only

- Find relevant apps in response to a search query

- Use semantic features to cluster similar apps(can be used for recommending apps or as a genre classifier)

Facebook App description

App Store Data

- Data source : Mobile App Statistics (Apple iOS app store)*

- 7200 iOS App store apps
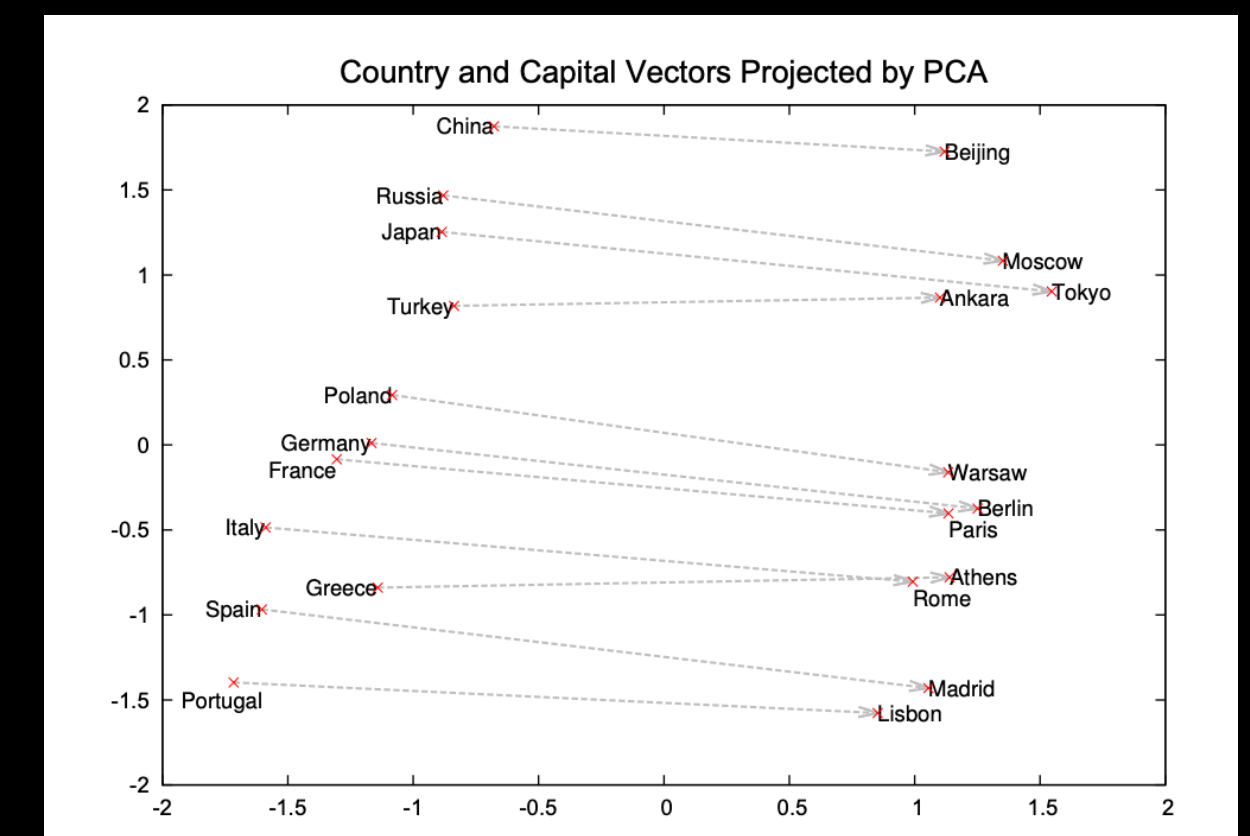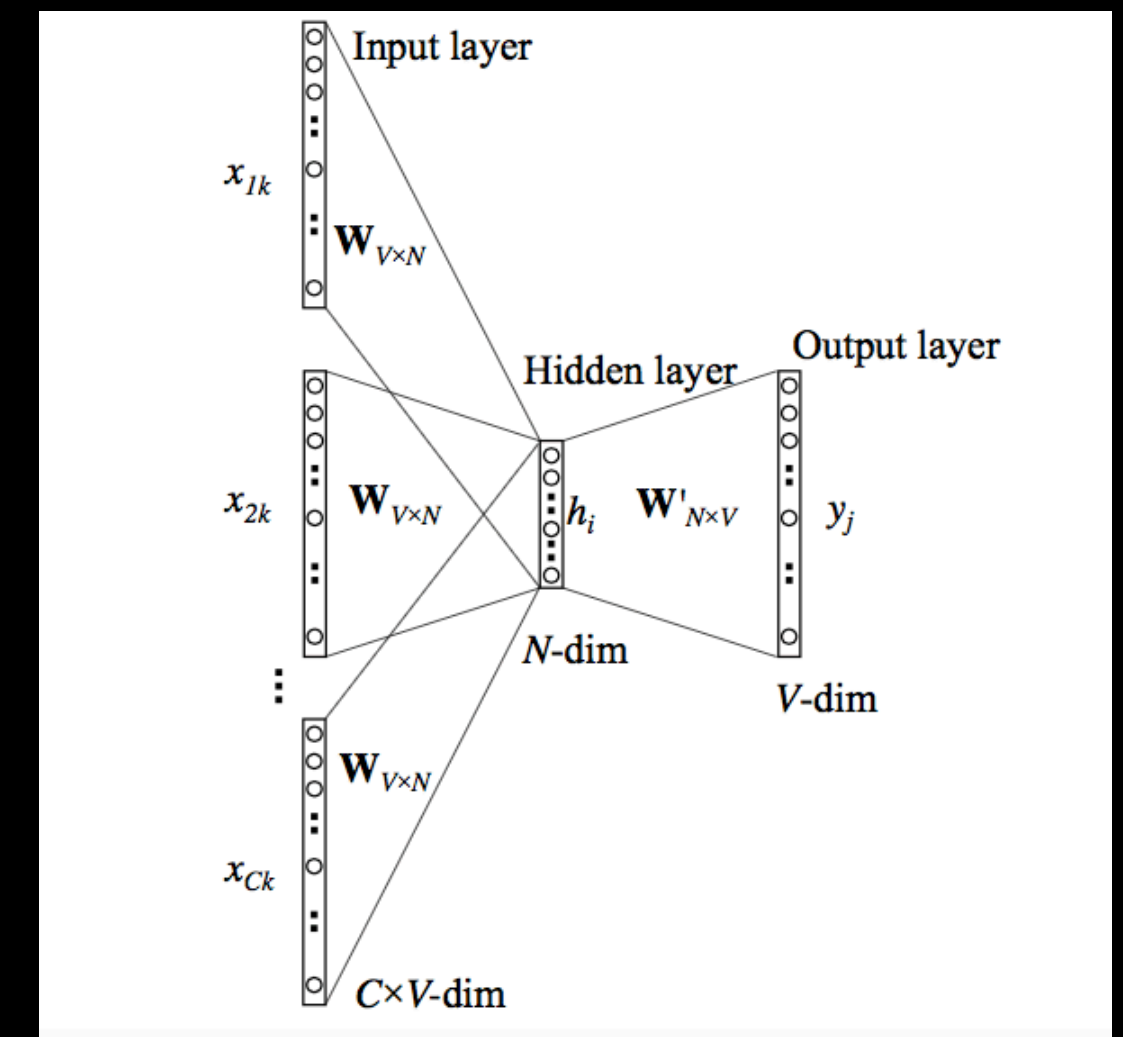
- Relevant fields: id, name, description

	apps_description			
	id	**track_name**	**size_bytes**	**app_desc**
0	281656475	PAC-MAN Premium	100788224	SAVE 20%, now only $3.99 for a limited time!\n...
1	281796108	Evernote - stay organized	158578688	Let Evernote change the way you organize your ...
2	281940292	WeatherBug - Local Weather, Radar, Maps, Alerts	100524032	Download the most popular free weather app pow...
3	282614216	eBay: Best App to Buy, Sell, Save! Online Shop...	128512000	The eBay app is the best way to find anything ...
4	282935706	Bible	92774400	On more than 250 million devices around the wo...
...	...		...	...	...
7192	1187617475	Kubik	126644224	Place the falling blocks correctly in order to...
7193	1187682390	VR Roller-Coaster	120760320	A thrilling virtual reality roller coaster exp...
7194	1187779532	Bret Michaels Emojis + Lyric Keyboard	111322112	Rock star Bret Michaels, winner of Celebrity A...
7195	1187838770	VR Roller Coaster World - Virtual Reality	97235968	VR Roller Coaster World is an app for Google C...
7196	1188375727	Escape the Sweet Shop Series	90898432	5 previous escape games plus 1 new game in one...

*https://www.kaggle.com/ramamet4/app-store-apple-data-set-10k-apps

Word2vec

Distributed representations of words in a vector space*

- Words represented by embeddings

- Helps achieve better performance in natural language processing tasks by grouping similar words.

- Trained on a large text corpus

- Vectors trained on part of Google News dataset (about 100 billion words)**.

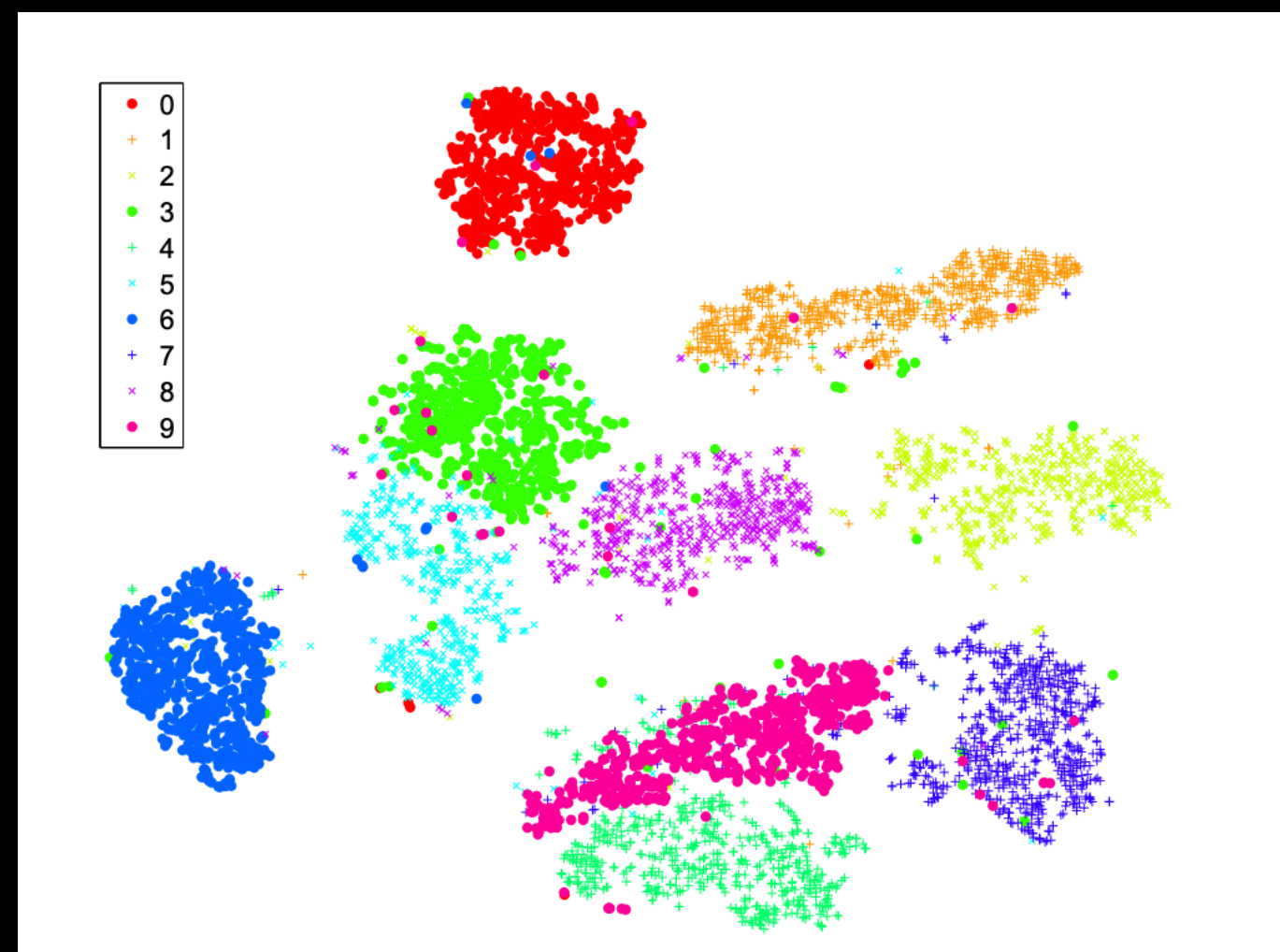- The model contains 300-dimensional vectors for 3 million words and phrases.

*https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

**https://code.google.com/archive/p/word2vec/

t-SNE

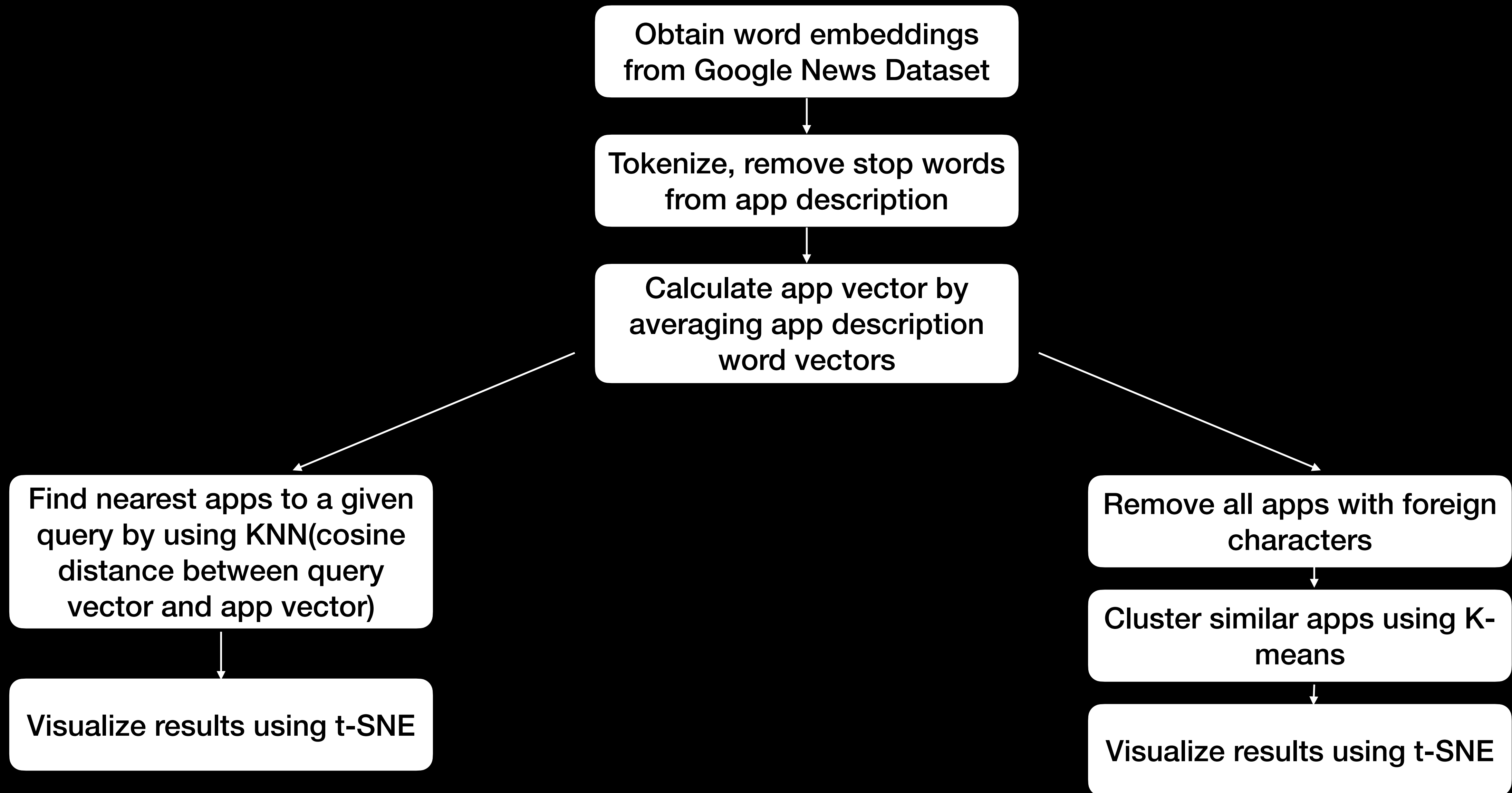**t-distributed stochastic neighbor embedding***

- Nonlinear dimensionality reduction technique

- For embedding high-dimensional data for visualization in a low-dimensional space

- Similar objects are assigned a higher probability while dissimilar points are assigned a very low probability.

Visualization by t-SNE on 6000 images from the MNIST dataset(784 dimensions)   *http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

Project Flowchart

Obtain word embeddings from Google News Dataset

Tokenize, remove stop words from app description

Calculate app vector by averaging app description word vectors

Find nearest apps to a given query by using KNN(cosine distance between query vector and app vector)

Visualize results using t-SNE

Remove all apps with foreign characters

Cluster similar apps using K-means

Visualize results using t-SNE

Results
Find relevant apps for a search query

```
['AccuWeather – Weather for Life',
 'WeatherPro',
 'MyRadar Pro NOAA Weather Radar, Forecasts & Storms',
 'Boating Weather',
 'WeatherBug – Local Weather, Radar, Maps, Alerts',
 'Living Earth – Clock & Weather',
 'The Weather Channel: Forecast, Radar & Alerts',
 'MyRadar NOAA Weather Radar Forecast',
 'NOAA Hi-Def Radar Pro –  Storm Warnings, Hurricane Tracker & Weather Forecast',
 'Virtual Regatta Offshore']
```
Weather

```
['Food Additives Checker (E Numbers)',
 'Sky Burger – Build & Match Food Free',
 'Allrecipes Dinner Spinner',
 'Paprika Recipe Manager',
 'Whole Foods Market',
 'Alipay – Makes Life Easy',
 'OpenTable – Restaurant Reservations',
 'Yelp – Nearby Restaurants, Shopping & Services',
 'Starbucks',
 "Sam's Club: Wholesale Shopping & Bulk Buy Deals"]
```
Food

```
['CNN: Breaking US & World News, Live Video',
 'CBS News – Watch Free Live Breaking News',
 'ABC News – US & World News + Live Video',
 'Fox News',
 'ESPN: Get scores, news, alerts & watch live sports',
 'Ameba',
 'HuffPost – News, Politics & Entertainment',
 'Twitter',
 'NBC News',
 'Sky News']
```
Media

```
['CBS Sports App – Sports Scores, News, Stats, Watch',
 'Yahoo Sports – Teams, Scores, News & Highlights',
 'FOX Sports Mobile',
 'Onefootball – Soccer Scores & Live News',
 'Yahoo Fantasy Sports',
 'Top Eleven 2017 – Be a Soccer Manager',
 'NFL',
 'ESPN: Get scores, news, alerts & watch live sports',
 'ABC News – US & World News + Live Video',
 "Ricky Carmichael's Motocross Matchup Pro"]
```
Sports

```
['Lose It! – Weight Loss Program and Calorie Counter',
 'Nike+ Training Club – Workouts & Fitness Plans',
 'Lifesum – Inspiring healthy lifestyle app',
 'Map My Ride+ – GPS Cycling & Route Tracker',
 'Sleep Cycle alarm clock',
 'Period Tracker Deluxe',
 'Map My Run+ – GPS Running & Workout Tracker',
 'Period Tracker Lite',
 'Weight Watchers',
 'SmartGo Player']
```
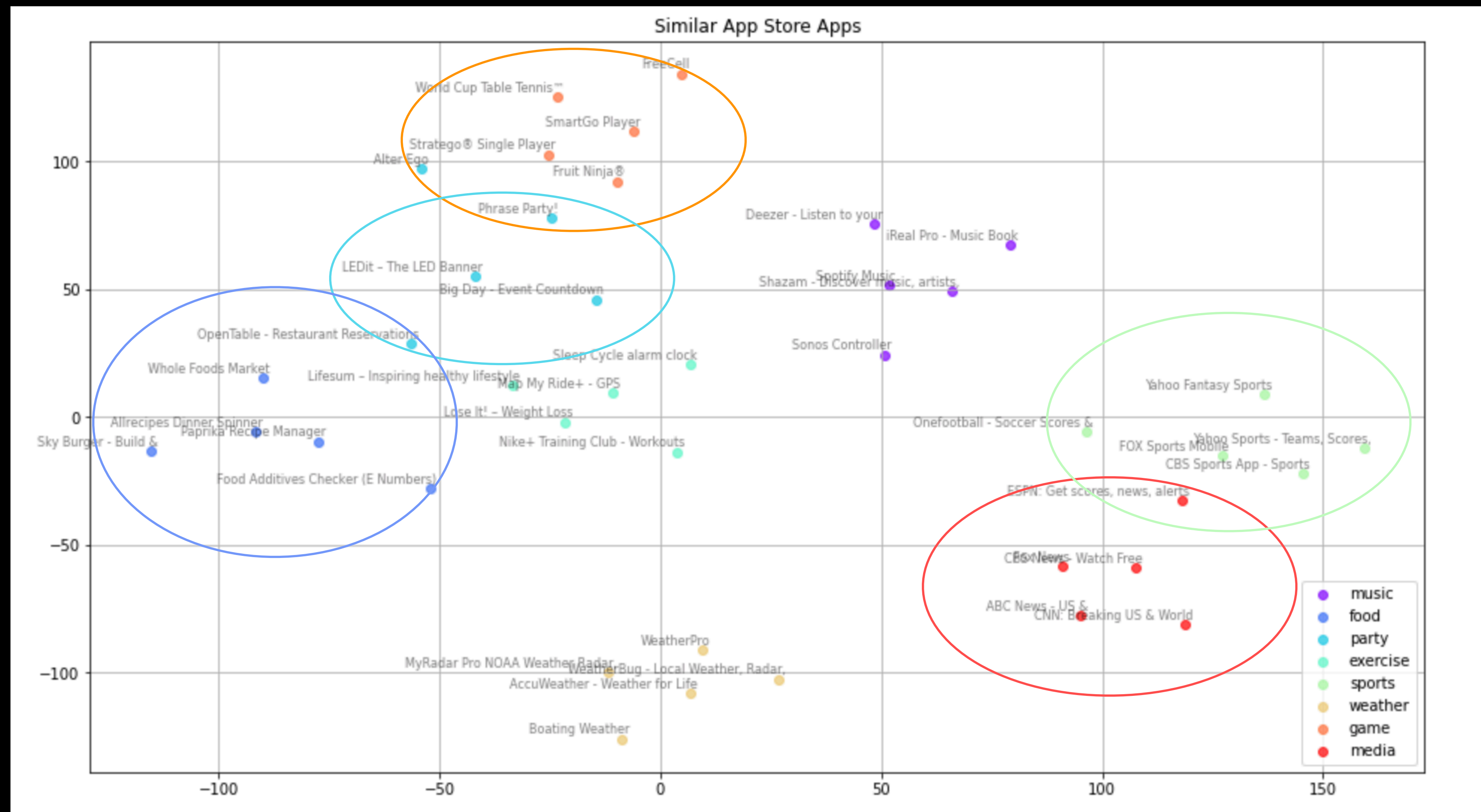Exercise

```
['Big Day – Event Countdown',
 'Phrase Party!',
 'LEDit – The LED Banner App',
 'Alter Ego',
 'OpenTable – Restaurant Reservations',
 'Wedding Dash Deluxe',
 'Alipay – Makes Life Easy',
 'Election 2016 Map',
 'Peekaboo Barn',
 'Virtual Families']
```
Party

Results
Visualizing similar apps using t-SNE

Similar App Store Apps

Results
Real life user queries

```
find_apps('mystery game',10)

['Canabalt',
 '▷Solitaire',
 'The Secret of Grisly Manor',
 'Neuroshima Hex',
 'Dungeon Raid',
 'LEGO Harry Potter: Years 1–4',
 'FreeCell',
 'Shoot The Zombirds',
 "Reiner Knizia's Ra",
 'Super Stickman Golf']
```

```
find_apps('social media',10)

['Twitter',
 'CBS News — Watch Free Live Breaking News',
 'CNN: Breaking US & World News, Live Video',
 'ABC News — US & World News + Live Video',
 'HuffPost — News, Politics & Entertainment',
 'NBC News',
 'ESPN: Get scores, news, alerts & watch live sports',
 'Sky News',
 'Parking Mania',
 'LinkedIn']
```

```
find_apps('watch tv',10)

['CNN: Breaking US & World News, Live Video',
 'ESPN: Get scores, news, alerts & watch live sports',
 'Twitter',
 'MSNBC',
 'Sunday Ticket',
 'NFL Sunday Ticket',
 'ABC News — US & World News + Live Video',
 'CBS News — Watch Free Live Breaking News',
 'The Masters Tournament',
 'DIRECTV']
```

```
find_apps('pay bills',10)

['Alipay — Makes Life Easy',
 'Mint: Personal Finance, Budget, Bills & Money',
 'PayPal — Send and request money safely',
 'iLoan Calc (Loan calculator)',
 'HomeBudget with Sync',
 'Dropbox',
 'Starbucks',
 'Bank of America — Mobile Banking',
 'Chase Mobile℠',
 'Zillow Real Estate — Homes for Sale & for Rent']
```

```
find_apps('watch football',10)

['CBS Sports App — Sports Scores, News, Stats, Watch',
 'Sunday Ticket',
 'NFL Sunday Ticket',
 'Yahoo Sports — Teams, Scores, News & Highlights',
 'FOX Sports Mobile',
 'Top Eleven 2017 — Be a Soccer Manager',
 'PGA Championship 2016 — Baltusrol Golf Club',
 'ESPN: Get scores, news, alerts & watch live sports',
 'NFL',
 'The Championships, Wimbledon 2016 — Tennis Grand Slam']
```

```
find_apps('wedding party',10)

['Wedding Dash Deluxe',
 'Big Day — Event Countdown',
 'Alter Ego',
 'OpenTable — Restaurant Reservations',
 'LEDit — The LED Banner App',
 'Shutterfly: Prints, Photo Books, Cards Made Easy',
 'Match™ — #1 Dating App.',
```

Results
Clustering similar apps

App Clusters

['Allrecipes Dinner Spinner',
 'Sky Burger —',
 'Cake Doodle',
 'Paprika Recipe Manager',
 "Annabel Karmel's Baby",
 "Domino's Pizza USA",
 'Green Kitchen —',
 'Toca Kitchen',
 'Nom Nom Paleo',
 'Cooking Academy (Full)']

['Google Translate',
 'Lieferando.de',
 'AliExpress Shopping App',
 'SayHi Translate',
 'ConjuVerb — Spanish',
 'Think in Japanese',
 'Translator !!',
 'iTranslate Voice —',
 'Voice Translate Pro',
 'HelloTalk Language Exchange']

["franceinfo — l'actualité",
 'AlloCiné : Cinéma',
 '6play, TV en',
 'Résultats Foot en',
 'Météo-France',
 'Ma Banque',
 'La Banque Postale',
 'Programme TV Télé',
 'Mon compte pour',
 'Impots.gouv']

['Color Splash',
 'iSwap Faces',
 'Shutterfly: Prints, Photo',
 'Camera+',
 'Camera Plus: For',
 'CatPaint',
 'Hipstamatic',
 'Timetracks — Slit-Scan',
 'Dr. Seuss Camera',
 'Slow Shutter Cam',
 'AgingBooth',
 'iWatermark — Watermark',
 'Drawing Pad',
 'Graphic — illustration']

['Line Rider iRide™',
 '2XL Supercross',
 'AirCoaster — Roller',
 'Drift Mania Championship',
 'Autohome—Find new&Used Cars',
 'Bloons Super Monkey',
 'Virtual Regatta Offshore',
 '2XL Supercross HD',
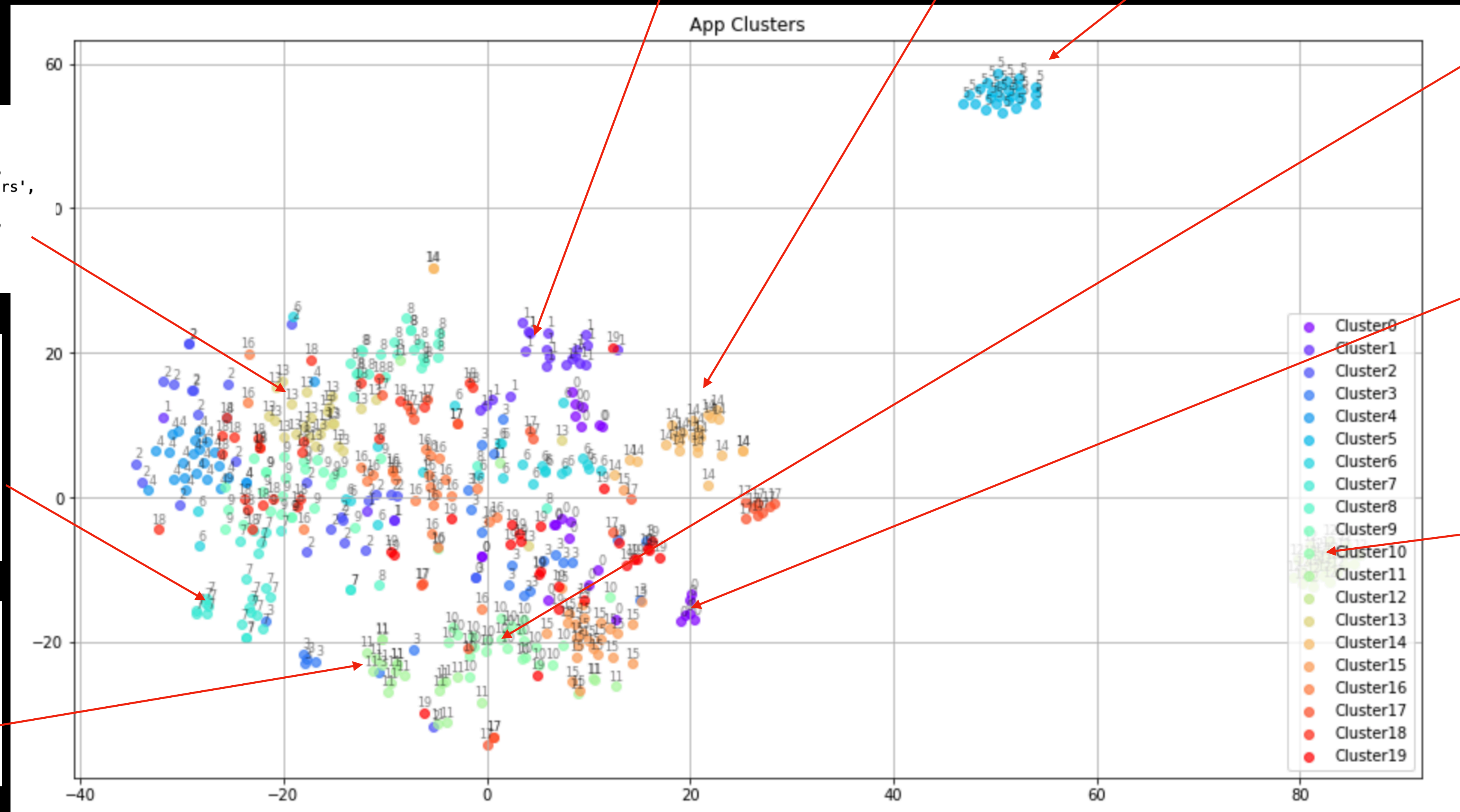 'Lane Splitter',
 'Mountain Biker']

['PayPal — Send',
 'Bank of America',
 'TripAdvisor Hotels Flights',
 'Yelp — Nearby',
 'Lifesum — Inspiring',
 'LinkedIn',
 'Deliveries: a package',
 'HotSchedules',
 'OpenTable — Restaurant',
 'Lose It! —',
 'Target',
 'Amazon App: shop,',
 'Chase Mobile℠',
 'Mint: Personal Finance,']

['Blackjack by MobilityWare',
 'FOX Sports Mobile',
 'Phase 10 Pro',
 'Beer Pong Game',
 'CBS Sports App',
 'The Masters Tournament',
 'World Cup Table',
 '2016 U.S. Open',
 'The Championships, Wimbledon',
 'Fantasy Football Manager',
 'PGA Championship 2016',
 'NFL Sunday Ticket',
 '2016 US Open',
 'Yahoo Fantasy Sports']

['radio.de — Der',
 'n-tv Nachrichten',
 'TV SPIELFILM —',
 'Jourist Weltübersetzer',
 'Der Feueralarm',
 'Tagesschau',
 'Verbrechen — echte',
 'Chefkoch — Rezepte,',
 'FOCUS Online —',
 'Oje, ich wachse!']

['Pandora — Music',
 'Shazam — Discover',
 'Cleartune — Chromatic',
 'White Noise',
 'Deezer — Listen',
 'Sonos Controller',
 'iReal Pro —',
 'Rowmote: Remote Control',
 'Tempo — Metronome',
 'Brain Wave ™']

Cluster0
Cluster1
Cluster2
Cluster3
Cluster4
Cluster5
Cluster6
Cluster7
Cluster8
Cluster9
Cluster10
Cluster11
Cluster12
Cluster13
Cluster14
Cluster15
Cluster16
Cluster17
Cluster18
Cluster19

Conclusions

- Word2vec gives good contextual information about app descriptions

- Search relevance can be improved by having more features like user intent, past download history, etc

- Training word2vec on App Store download data can be even more accurate in getting semantic features

Thank you