Content Based Recommendation System

Business Driver

- Recommendations generate 35% of Amazon's revenue
- "Improves with use" - Better customer retention
- Increases customer basket size
- Personalization improves customer loyalty
- Cross sell opportunity

Content Based Recommendation

Product Features

- Title – Brief description of the product
- Product Type – Shirt, Tshirt, Pants
- Color
- Brand
- Image URL

Nearly duplicate Titles

100% Cotton Camisole 4-Pack, Assorted, Size Extra Large (2X)

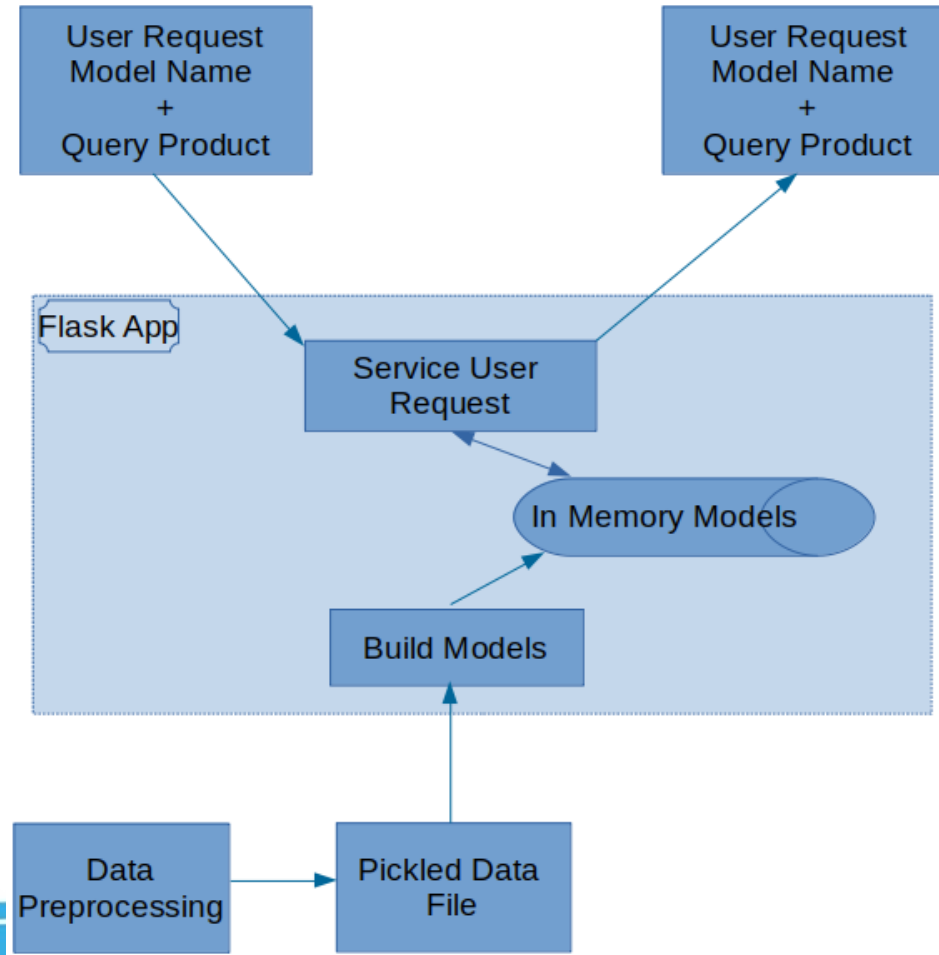100% Cotton Camisole 4-Pack, Size Extra Large (3X)

100% Cotton Camisole 4-Pack, Size Small (3X)

High cosine similarity

But all results show the same image

Workflow

Models

- Bag of Words
- TfIdf
- Idf
- Average Word2Vec
- Idf Weighted Word2Vec
- Word2Vec + Brand

Average Word2Vec

- Use the 300 dimension Google News Word2Vector

- Sum up all the vectors of the words in Title and compute average

- Average Vector of similar products are closer to one another

Title Text	300 Dimension Word2Vec Vector			
women	1.2	2.2	1.7	1.1
long	0.1	2.6	1.2	1.3
sleeve	2.1	2	1.3	1.1
Average Word2Vec	**0.73**	**2.3**	**1.4**	**1.2**

Idf Weighted Word2Vec

- Each word vector is multiplied by its IDF value

- IDF Model provides the IDF value for the word

- Word2Vector * IDF Value

Title Text	Word2Vec Vector				X	IDF Value
women	1.2	2.2	1.7	1.1	X	0.4
long	0.1	2.6	1.2	1.3	X	0.6
sleeve	2.1	2	1.3	1.1	X	0.3
Result						
women	0.48	0.9	0.8	0.44		
long	0.06	1.5	1.2	0.72		
sleeve	0.63	0.6	0.4	0.3		
Average Word2Vec	**1.17**	**0.9**	**0.8**	**0.48**		

Word2Vec + Brand

Title Word Vec = w1 * Word2Vec Vector + w2 * Brand Vector

Title Text	Word2Vec Vector					Brand		
women	1.2	2.2	1.7	1.1				
long	0.1	2.6	1.2	1.3				
sleeve	2.1	2	1.3	1.1				
Average Word2Vec	**1.13**	**2.3**	**1.4**	**1.16**		0	1	1

Thank you!