

# DUNNHUMBY SHOPPERS' MISSIONS

## STAT 645: Mini-Project 2

---

Joshua Fagan, Alex Klibisz

October 27, 2016

University of Tennessee

Introduction

Data Review

Data Preparation

LDA with Low-Entropy Naming

AT with Low-Entropy Naming

Future Work

# INTRODUCTION

---

1. Dunnhumby Dataset.
2. Identify latent missions for shoppers.
3. Link missions to demographics and/or product categories.

1. Can we group products into distinguishable categories?
2. Can we identify demographic attributes that distinguish specific shoppers from the rest of the population?
3. Can we identify a temporal pattern in shoppers' topics?
4. Can we show how the distinguishing attributes can be leveraged for more effective marketing campaigns?

# DATA REVIEW

---

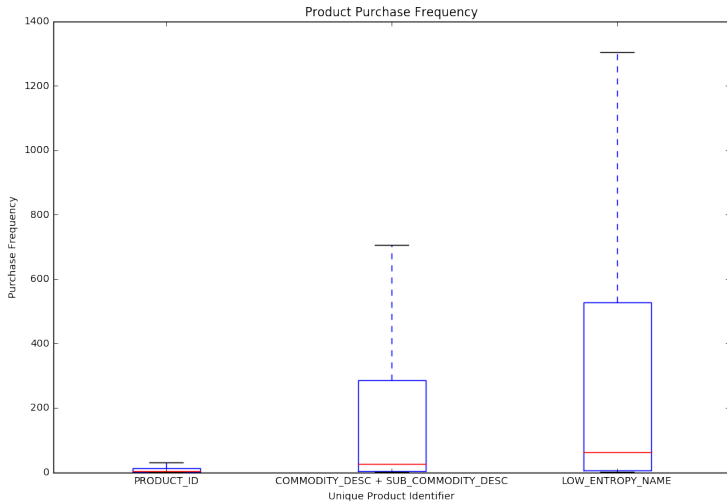
## High Level Numbers

- 2,595,732 unique transactions
- 276,484 unique shopping baskets
- 92,339 unique products
- 2,500 unique households

## Mapping to Topic Model

- Corpus is all baskets across all transactions.
- Documents are baskets.
- Words are products in each basket.
- (Products can be identified in various ways.)
- (Authors can be identified based on household key or demographic classifications.)

# IDENTIFYING UNIQUE PRODUCTS



**Figure:** If you use PRODUCT\_ID as an identifier, the same product gets purchased relatively few times



# INFLATED GASOLINE QUANTITIES

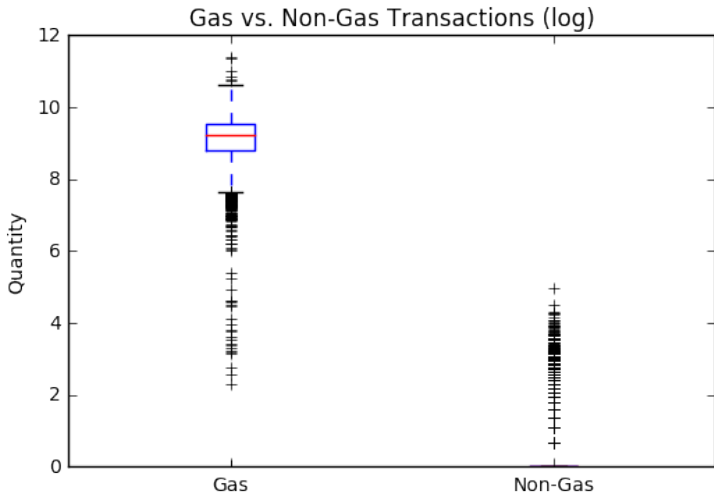


Figure: Mean quantity for GASOLINE-REG UNLEADED was 10,308!

# DATA PREPARATION

---

## Goals

1. Avoid similarly named products in different categories.
2. Similarly-named products should be identified as the same product.

## Method

1. Tokenize every product: Combine COMMODITY\_DESC + SUB\_COMMODITY\_DESC, split into words, remove punctuation, stem, remove stop-words, replace synonyms.
2. Count frequency of tokens across corpus.
3. Each product's LOW\_FREQUENCY\_NAME is the product's two most frequent tokens.

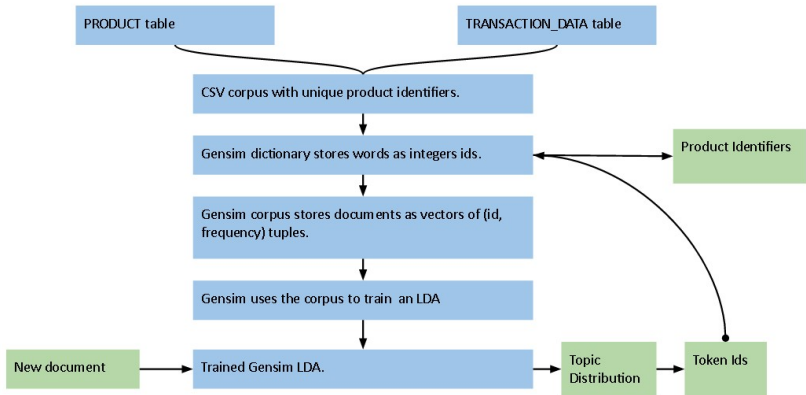
## LOW-ENTROPY NAMES

Low Entropy Name	Matching	Original Names
everyday spli	1323	GREETING CARDS/WRAP/PARTY SPLY PARTY EVERYDAY, GREETING CARDS/WRAP/PARTY SPLY CARDS EVERY- DAY
chocol candi	1016	CANDY - PACKAGED CANDY BAGS- CHOCOCLATE, CANDY - PACKAGED CANDY BOXED CHOCOLATES, NO COMMODITY DESCRIPTION CANDY BAGS-CHOCOCLATE, CANDY - PACKAGED SEASONAL CANDY BAGS-CHOCOLATE, CANDY - PACKAGED SEASONAL CANDY BOX-CHOCOLATE W, CANDY - PACKAGED SEASONAL CANDY BOX-CHOCOLATE
beer ale	772	BEERS/ALES BEERALEMALT LIQUORS
entre premium	665	FRZN MEAT/MEAT DINNERS FRZN SS PREMIUM ENTREES/DNRS/T, FRZN MEAT/MEAT DINNERS FRZN SS PREMIUM

# LDA WITH LOW-ENTROPY NAMING

---

# LDA PIPELINE



**Figure:** Pipeline for preparing documents, fitting the LDA, and evaluating new documents.

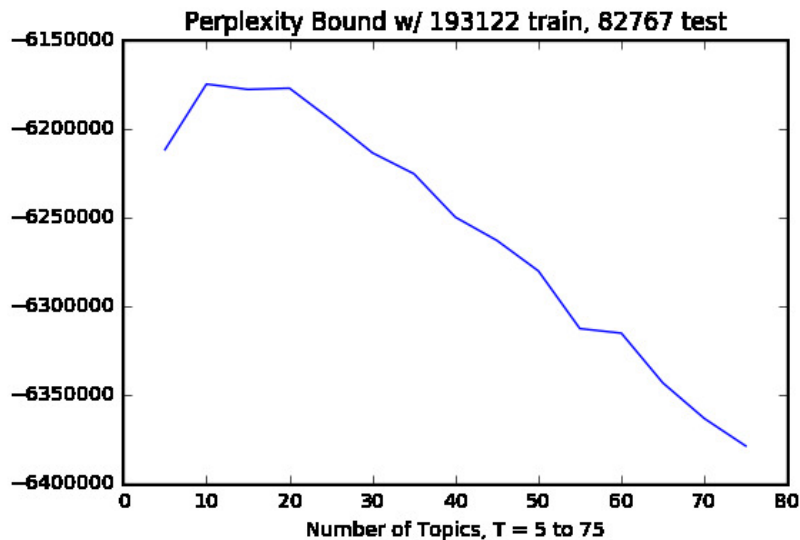


Figure: LDA perplexity lower bound; note that higher is better.

Topic 0 Word	Low-Entropy Word	$p(w z)$
SOFT DRINKS 12/18&15PK CAN CAR	15pk car	0.29
GASOLINE-REG UNLEADED	unlead fuel	0.14
BEERALEMALT LIQUORS	beer ale	0.09
SW GDS:DONUTS	donut breakfast	0.06
CARDS SEASONAL	spli greet	0.02
PREP FD: SIDE DISHES (HOT)	prep side	0.02
COLD AND FLU - DXM	dxm flu	0.02
TORTILLA/NACHO CHIPS	nacho tortilla	0.01
ROLLS: DINNER	roll dinner	0.01
BREAD:ITALIAN/FRENCH	italian french	0.01

**Table:** LDA top ten words in topic 0 ( $T = 10$ ), "convenience store topic"



Topic 5 Word	Low-Entropy Word	$p(w z)$
FLUID MILK WHITE ONLY	white fluid	0.09
BANANAS	banana tropic	0.03
ORANGES NAVELS ALL	navel orang	0.02
CIGARETTES	cigarett	0.02
DAIRY CASE 100% PURE JUICE - O	pure 100	0.02
EGGS - X-LARGE	egg	0.01
ALL FAMILY CEREAL	famili cereal	0.01
MAINSTREAM	wheat multigrain	0.01
WHEAT/MULTIGRAIN BR		
CREAM CHEESE	cream chees	0.01
MAINSTREAM WHITE BREAD	white bun	0.01

**Table:** LDA top ten words in topic 5 (T = 10)

Topic 9 Word	Low-Entropy Word	$p(w z)$
BABY FOOD - BEGINNER	beginn babi	0.06
PAPER TOWELS & HOLDERS	holder towel	0.04
BABY FOOD JUNIOR ALL BRANDS	junior babi	0.03
CANDY BARS (MULTI PACK)	multi pack	0.03
TOILET TISSUE	toilet tissu	0.03
FACIAL TISSUE & PAPER HANDKE	handk tiss	0.03
CHEWING GUM	gum chew	0.03
LIQUID LAUNDRY DETERGENTS	deterg laundri	0.02
BABY DIAPERS	diaper dispos	0.02
BAR SOAP	soap liquid	0.02

**Table:** LDA top ten words in topic 9 (T = 10), "parent topic"

## Intuition

- Given a household falls into  $\text{INCOME\_DESC} = X$ , what is the probability of purchasing a product from each of the topics?
- How **different** is that demographic distribution relative the population?

## Method

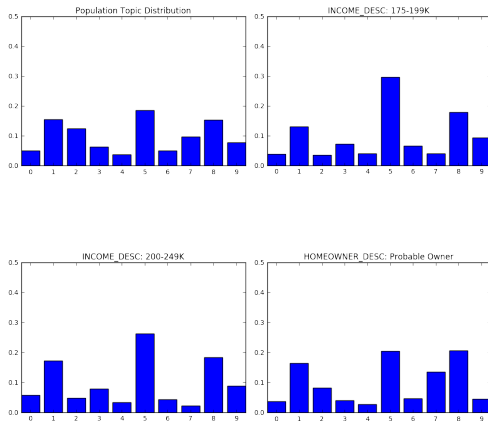
- Concatenate entire corpus to compute population distribution.
- Narrow down specific demographics, concatenate their documents to compute demographic distributions.
- Use KL-divergence to quantify difference.

## DEMOGRAPHIC DISTRIBUTIONS

	Demographic	Category	Entropy	Top Topic
0	INCOME_DESC	175-199K	0.102886	5
1	INCOME_DESC	200-249K	0.093301	5
2	INCOME_DESC	250K+	0.060177	5
3	INCOME_DESC	150-174K	0.052650	5
4	INCOME_DESC	100-124K	0.046601	5
5	HOMEOWNER_DESC	Probable Owner	0.040425	8
6	INCOME_DESC	125-149K	0.033437	5
7	AGE_DESC	19-24	0.027506	1
8	AGE_DESC	65+	0.024285	5
9	HH_COMP_DESC	1 Adult Kids	0.020416	8
10	INCOME_DESC	Under 15K	0.019510	8

**Table:** Top 10 entropy values for demographic distributions over topics relative the population distribution

# DEMOGRAPHIC DISTRIBUTIONS



**Figure:** LDA distribution of all documents over topics and the 1st, 2nd, and 5th most different demographic distributions.

# AT WITH LOW-ENTROPY NAMING

---

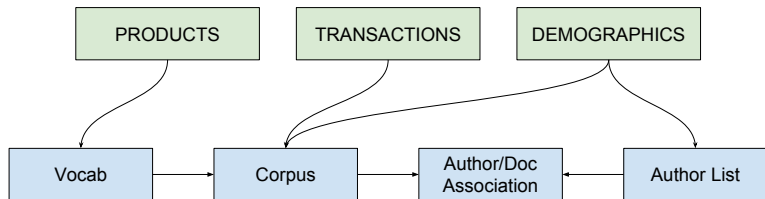


Figure: Pipeline for preparing documents and fitting the AT.

- Number of documents: 140,339
- Number of unique words: 18,620
- 41 Demographic Authors
- 801 Household Key Authors
- Topics: 10, 15, 20, 25, 30, 40, 50
- Gibbs Sampling Iterations 10, 50



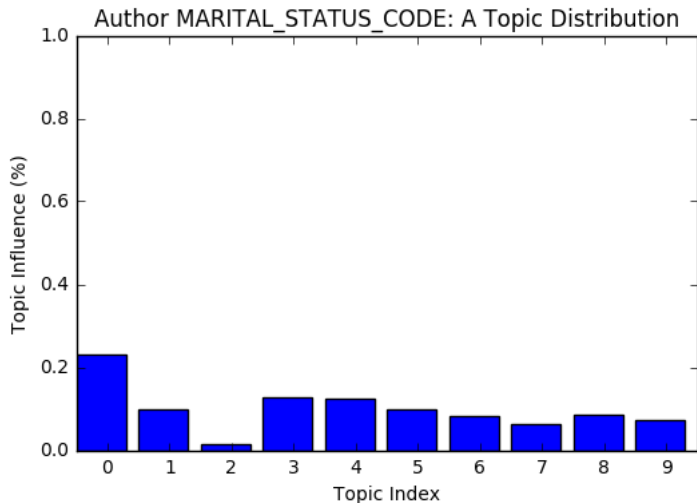
Low-Entropy Word	$p(w z)$
incl liter: 2 LTR	0.0435400775017
white bun: 20 OZ	0.0194244916306
ramen cup: 3 OZ	0.0168338102399
srv sngl: 20 OZ	0.0165095036903
pasta stabl: 15 OZ	0.0142959229387
15pk car: 12 OZ	0.013945219344s
lean beef:	0.0110867964999
cigarett: CTN	0.010113876851
economi entre: 8 OZ	0.0100610827615
white fluid:	0.00906176606788

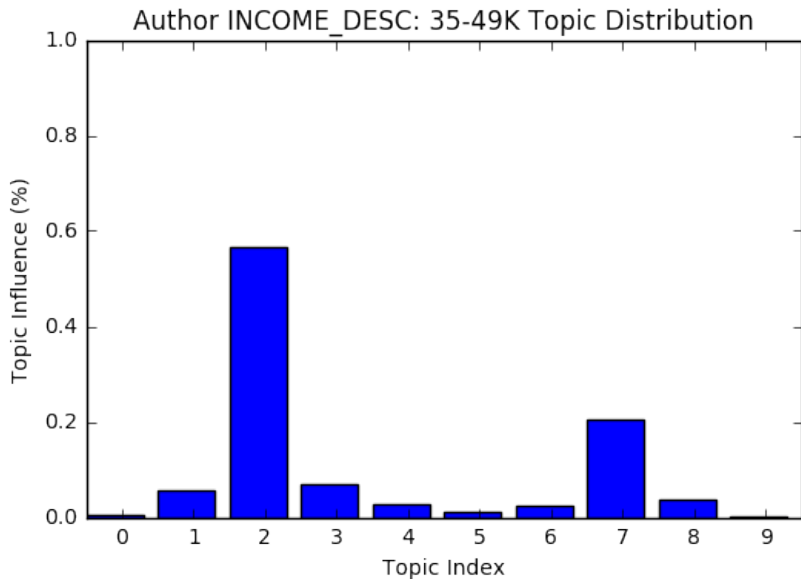
**Table:** Topic 2 top ten words

Low-Entropy Word	$p(w z)$
not yogurt: 6 OZ	0.0556582883036
strawberri berri: 16 OZ	0.0160711494707
lemon citru:	0.0138042548419
banana tropic: 40 LB	0.0128796004538
bun roll: 24 OZ	0.0128080143077
lime citru: 36CT	0.011400153433
italian french:	0.0112092570432
crown cauliflow:	0.0103740853378
boneless breast:	0.00962243080304
homestyl chunki: 18.8 OZ	0.00896622446313

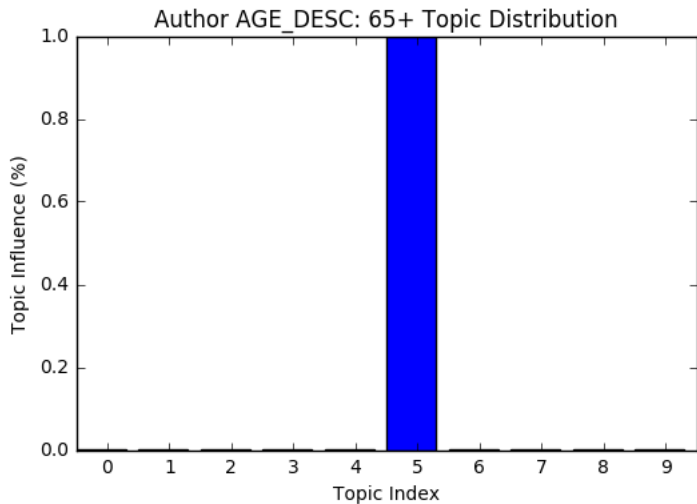
**Table:** Topic 9 top ten words

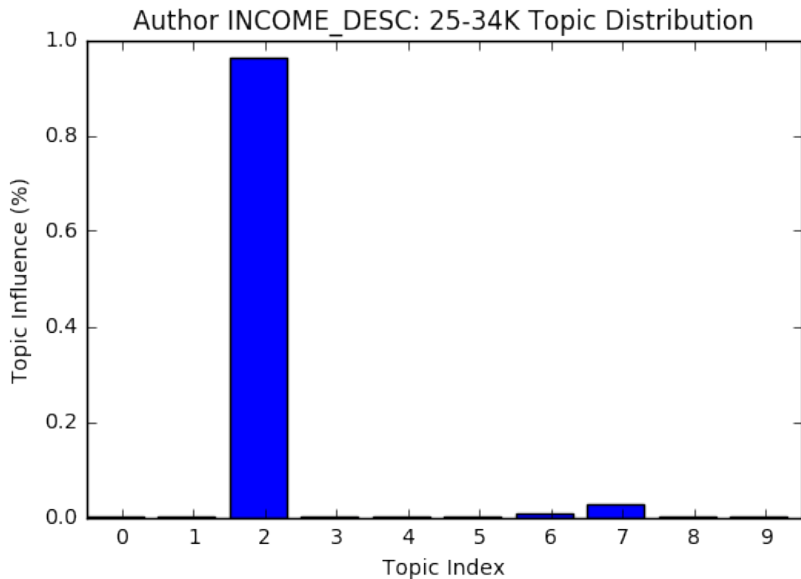
## SINGLE AUTHOR TOPIC DISTRIBUTIONS



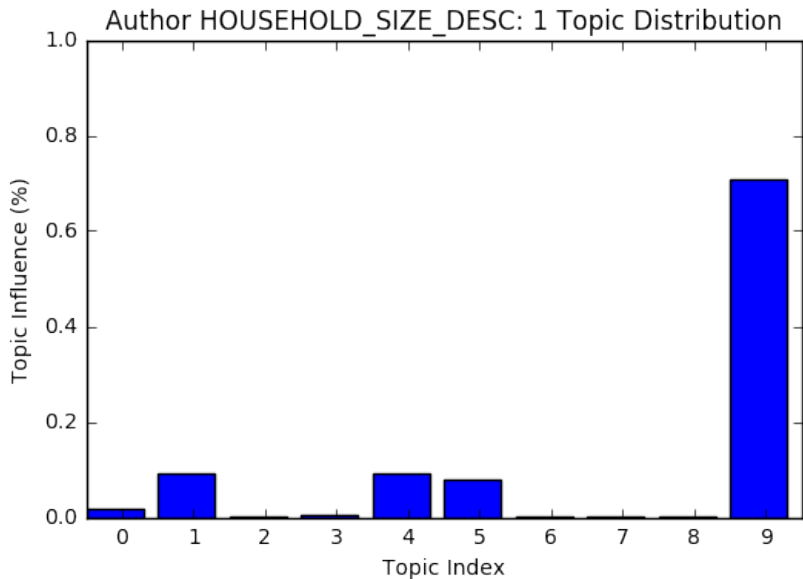


## SINGLE AUTHOR TOPIC DISTRIBUTIONS

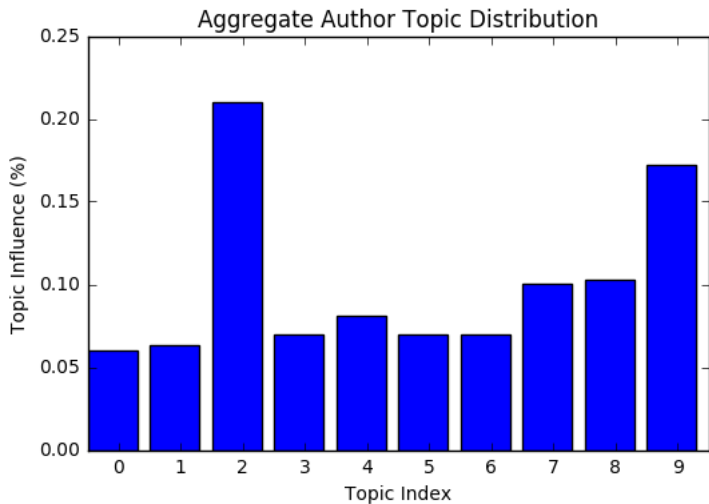




## SINGLE AUTHOR TOPIC DISTRIBUTIONS



# AGGREGATE AUTHOR DISTRIBUTION



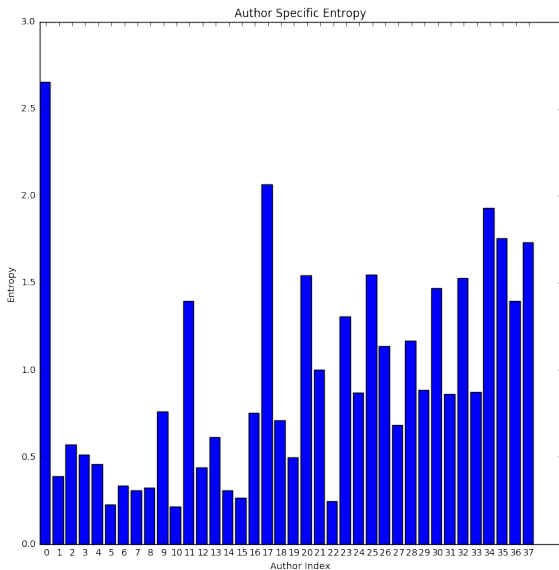


---

Author	Entropy
AGE_DESC: 65+	2.65500
HOUSEHOLD_SIZE_DESC: 4	2.064694
AGE_DESC: 19-24	1.928755
AGE_DESC: 55-64	1.753791
HOMEOWNER_DESC: Probable Owner	1.731984
INCOME_DESC: 15-24K	1.548079
HH_COMP_DESC: Single Female	1.544354
KID_CATEGORY_DESC: 3+	1.525487
HH_COMP_DESC: Single Male	1.469682
INCOME_DESC: 25-34K	1.397155

---

# AGGREGATE AUTHOR DISTRIBUTION



# FUTURE WORK

---

- Model incorporating temporal data - HMM, LSTM
- Question: Does a demographic's topical preference change over time?
- Question: Do coupons and advertising campaigns create a change in topical interest?