Project 4: NLP: Google and Apple Sentiment Analysis

Part 2: Post Modeling EDA and Data Viz

This is for additional things like LDA, visualizations, and Neural Nets (aka Deeper NLP...)
Also includes some basic comparative examination of actual tweets to better understand the outputs of the LDA process.

Topic Modeling with LDA

Following instructions / code from: Topic Modeling in Python: Latent Dirichlet Allocation (LDA) - towarddatascience.com

In [1]:
```
# Load some relevant libraries.
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
import seaborn as sns
```

In [2]:
```
# Load the pre-saved data set
data = pd.read_csv('df5_tweets.csv')
data.head()
```

Out[2]:

	Unnamed: 0	text	brand	feelings	text_fixed	char_count	stopwords	polarity	subjectivity
0	0	.@wesley83 I have a 3G iPhone. After 3 hrs twe...	Apple	0	wesley83 3g iphone 3 hrs tweeting rise_austin ...	117	10	-0.200000	0.400000
1	1	@jessedee Know about @fludapp ? Awesome iPad/i...	Apple	1	jessedee know fludapp awesome ipadiphone app y...	130	5	0.466667	0.933333
2	2	@swonderlin Can not wait for #iPad 2 also. The...	Apple	1	swonderlin wait ipad 2 also sale	74	8	0.000000	0.000000
3	3	@sxsw I hope this year's festival isn't as cra...	Apple	0	hope years festival isnt crashy years iphone app	76	5	0.000000	0.000000
4	4	@sxtxstate great stuff on Fri #SXSW: Marissa M...	Google	1	sxtxstate great stuff fri marissa mayer google...	117	1	0.800000	0.750000

In [3]:
```
# Remove unneeded columns
data2 = data.drop(columns=['Unnamed: 0', 'char_count', 'stopwords', 'polarity', 'subjectivity'], axis=1)
data2.head()
```

Out[3]:

	text	brand	feelings	text_fixed
0	.@wesley83 I have a 3G iPhone. After 3 hrs twe...	Apple	0	wesley83 3g iphone 3 hrs tweeting rise_austin ...
1	@jessedee Know about @fludapp ? Awesome iPad/i...	Apple	1	jessedee know fludapp awesome ipadiphone app y...
2	@swonderlin Can not wait for #iPad 2 also. The...	Apple	1	swonderlin wait ipad 2 also sale
3	@sxsw I hope this year's festival isn't as cra...	Apple	0	hope years festival isnt crashy years iphone app
4	@sxtxstate great stuff on Fri #SXSW: Marissa M...	Google	1	sxtxstate great stuff fri marissa mayer google...

In [4]:
```
# Try this LDA stuff from scratch... get rid of text_fixed from previous EDA.
data2 = data2.drop(columns=['text_fixed'], axis=1)
data2.head()
```

Out[4]:

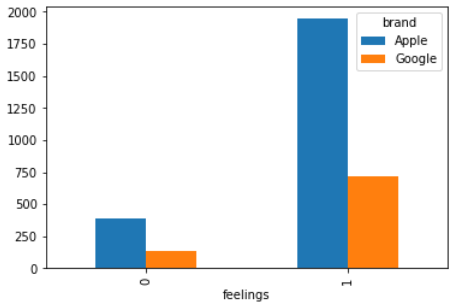
	text	brand	feelings
0	.@wesley83 I have a 3G iPhone. After 3 hrs twe...	Apple	0
1	@jessedee Know about @fludapp ? Awesome iPad/i...	Apple	1
2	@swonderlin Can not wait for #iPad 2 also. The...	Apple	1
3	@sxsw I hope this year's festival isn't as cra...	Apple	0
4	@sxtxstate great stuff on Fri #SXSW: Marissa M...	Google	1

n [27]:
```
data2.shape
```

ut[27]: (3182, 4)

n [99]:
```
# Recall that we do have some class imbalance (0 = negative tweets; 1 = positive tweets), many more positive than negative.
pd.crosstab(data2['feelings'],data2['brand']).plot.bar()
```

ut[99]: <AxesSubplot:xlabel='feelings'>



[124]:
```
# NOTE that from the Random Forest mpodel we were able to extract the feature importances... which shows us the top words ('tokens')
# contributing to the model. Additional EDA is required to help us better understand the context and meaning of these words.
# The top 10 words included: fail, headache, battery, long, need, think, design, suck, people, yet.
```

In [ ]:

Time for some post modeling EDA on the text (tweets) themselves...

After running the first LDA model (with all tweets) it became apparent more meaning could be gleaned by breaking up the tweets into a coupl different sets. So lets break it into a couple of different df.

n [29]:
```
# Consider splitting up into 4 dfs here...Apple positive; Google positive; Apple negative; Google negative
data3 = data2.loc[(data2['brand'] == 'Apple') & (data2['feelings'] == 1)]
print(data3.shape)
data3.head()
```

```
(1945, 4)
```

ut[29]:

	text	brand	feelings	text_processed
1	@jessedee Know about @fludapp ? Awesome iPad/i...	Apple	1	@jessedee know about @fludapp awesome ipad/ip...
2	@swonderlin Can not wait for #iPad 2 also. The...	Apple	1	@swonderlin can not wait for #ipad 2 also they...
6	Beautifully smart and simple idea RT @madebyma...	Apple	1	beautifully smart and simple idea rt @madebyma...
7	Counting down the days to #sxsw plus strong Ca...	Apple	1	counting down the days to #sxsw plus strong ca...
12	Great #sxsw ipad app from @madebymany: http://...	Apple	1	great #sxsw ipad app from @madebymany: http://...

n [44]:
```
data3.text[1:25]
```

ut[44]:
```
2     @swonderlin Can not wait for #iPad 2 also. They should sale them down at #SXSW.
6     Beautifully smart and simple idea RT @madebymany @thenextweb wrote about our #hollergram iPad app for #sxsw! http://bit.ly/ieaVOB (ht
7     Counting down the days to #sxsw plus strong Canadian dollar means stock up on Apple gear
12    Great #sxsw ipad app from @madebymany: http://tinyurl.com/4nqv92l (http://tinyurl.com/4nqv92l)
13    haha, awesomely rad iPad app by @madebymany http://bit.ly/hTdFim (http://bit.ly/hTdFim) #hollergram #sxsw
15    Just added my #SXSW flights to @planely. Matching people on planes/airports. Also downloaded the @KLM iPhone app, nicely done.
16    Must have #SXSW app! RT @malbonster: Lovely review from Forbes for our SXSW iPad app Holler Gram - http://t.co/g4GZypV (http://t.co/g
17    Need to buy an iPad2 while I'm in Austin at #sxsw. Not sure if I'll need to Q up at an Austin Apple store?
18    Oh. My. God. The #SXSW app for iPad is pure, unadulterated awesome. It's easier to browse events on iPad than on the website!!!
20    Photo: Just installed the #SXSW iPhone app, which is really nice! http://tumblr.com/x6t1pi6av7 (http://tumblr.com/x6t1pi6av7)
22    RT @LaurieShook: I'm looking forward to the #SMCDallas pre #SXSW party Wed., and hoping I'll win an #iPad resulting from my shameless
MC
23    RT haha, awesomely rad iPad app by @madebymany http://bit.ly/hTdFim (http://bit.ly/hTdFim) #hollergram #sxsw (via @michaelpiliero)
25    The new #4sq3 looks like it is going to rock. Update for iPhone and Android should push tonight http://bit.ly/etsbZk (http://bit.ly/e
stinWeird
27    Very smart from @madebymany #hollergram iPad app for #sxsw! http://t.co/A3xvWc6 (http://t.co/A3xvWc6) (may leave my vuvuzela at home
28    You must have this app for your iPad if you are going to #SXSW http://itunes.apple.com/us/app/holler-gram/id420666439?mt=8 (http://it
p/holler-gram/id420666439?mt=8) #hollergram
29    The best!  RT @mention Ha! First in line for #ipad2 at #sxsw &quot;pop-up&quot; Apple store was an event planner #eventprofs #pcma #e
31    @mention  - Great weather to greet you for #sxsw! Still need a sweater at night..Apple putting up &quot;flash store&quot; downtown to
32    #IPad2 's  Û÷#SmartCover  Ûª Opens to Instant Access - I should have waited to get one! - {link} #apple #SXSW
33    HOORAY RT  ÛÏ@mention Apple Is Opening A Pop-Up Store In Austin For #SXSW | @mention {link}
34    wooooo!!!  ÛÏ@mention Apple store downtown Austin open til Midnight. #sxsw  Û
36    {link} RT @mention 1st stop on the #SXSW #Chaos &amp; @mention hunt: Austin Java. Get in the spy game 4 a chance 2 win an iPad!
37    #OMFG! RT @mention Heard about Apple's pop-up store in downtown Austin? Pics are already on Gowalla: {link} #sxsw #iPad2
39    Check out @mention @mention &amp; @mention in line for their iPad 2 in Austin. Power to them! #sxswi #SXSW  {link}
42    I love my @mention iPhone case from #Sxsw but I can't get my phone out of it #fail
Name: text, dtype: object
```

n [73]:
```
data3.text.sample(20)
```

ut[73]:
```
754     Does the apple pop-up store still have iPads? #sxsw #sxswi
1914    RT @mention @mention just saw you at lax I'm heading to #Sxsw  If you or your wife have an iPhone or iPad download my app #Freespee
2862    Yo this #SXSW iPhone app is illa-def! Go!
974     Sweet! The convore iPhone app is ready for #sxsw : {link}
535     Very wise... RT @mention Apple is opening up a temp store in Austin for #SXSW {link} via @mention #iPad2
2273    RT @mention New #UberSocial for #iPhone now in the App Store includes UberGuide to #SXSW  {link}
962     Looks like all the apps for #SXSW are for the #iPhone. What about the #blackberry? The blackberry??? {link} (rt @mention
946     Apple is driving the &quot;consumerization&quot; of IT.  #empowered  #SXSW
3168    @mention you are my favorite-- thanks for coming to @mention -- when you getting an iPhone again?? #SXSW
2925    I'm a captain penguin now! RT @mention congrats to @mention for getting to the next level in his fave iPad game PengAirborne #SXSW
593     The only #airline to be mentioned by Guy Kawasaki as delightful like Apple (drumroll)... @mention (!!) Congrats! #sxsw
202     Before It Even Begins, Apple Wins SXSW {link} #Apple #iPad2 #PM #SXSW
2671    &quot;O frabjous day ! Callooh ! Callay !  Û_[Baaah!?]&quot;  ÛÓLewis\n Carroll   ÷È crowd swarms for iPad 2 launch {link} via @ment
195     Before It Even Begins, Apple Wins #SXSW {link} /by @mention for @mention
2426    RT @mention w00t! @mention wrote about our #hollergram iPad app for #sxsw! {link}
267     For those in need of sweet Mac goodness at #SXSW, Apple has set up a temporary store downtown: {link}
2373    RT @mention Sweet, Apple's opening a pop-up shop in the Scarbrough Building on Congress for the iPad 2 - {link} /via @mention #sxsw
1558    Apple selling iPad 2 at #sxsw. California is great at building cult followings. {link}
1770    iPad 2. Another reason I'd love to be @mention #sxsw
996     one of the most in-your-face ex. of stealing the show in yrs RT @mention &quot;At #SXSW, Apple schools the mkt experts&quot;  {link
Name: text, dtype: object
```

SUMMARY of Apple positive tweets

```
iPad2 excitement;building cult following;
purchasing apple proeducts / gear ; SmartCover; iPhone case
purchase at Austin apple store / flash store / pop-up store

iPad app for conference - madebymany - Holler Gram - ease of use
planely app - flights, planes, airports
convore app?
4Square app is better

Apple as driving consumerization of IT; building cult following; beating the mrkt. experts

party!
```

n [32]:
```
data2.shape
```

ut[32]:
```
(3182, 4)
```

[123]:
```
# data4.text[1:25]
```

n [33]:
```
data4 = data2.loc[(data2['brand'] == 'Google') & (data2['feelings'] == 1)]
print(data4.shape)
data4.head()
```

```
(719, 4)
```

ut[33]:

	text	brand	feelings	text_processed
4	@sxtxstate great stuff on Fri #SXSW: Marissa M...	Google	1	@sxtxstate great stuff on fri #sxsw: marissa m...
5	#SXSW is just starting, #CTIA is around the co...	Google	1	#sxsw is just starting #ctia is around the cor...
8	Excited to meet the @samsungmobileus at #sxsw ...	Google	1	excited to meet the @samsungmobileus at #sxsw ...
9	Find & Start Impromptu Parties at #SXSW Wi...	Google	1	find & start impromptu parties at #sxsw wi...
10	Foursquare ups the game, just in time for #SXS...	Google	1	foursquare ups the game just in time for #sxsw...

n [74]:
```
data4.text.sample(20)
```

ut[74]:
```
1221    &quot;Google before you tweet&quot; is the new &quot;think before you speak.&quot; - Mark Belinsky, #911tweets panel at #SXSW.
2818    At #SXSW seeing a demo of #Google maps for mobile 5.2. 3D rotational viewing is very cool
1137    Mayer comes out sans intro, still gets cheers. #techrockstar Launches into Google's priority on location - Fast, Fun &amp; Future #
2139    RT @mention Google's Marissa Mayer on the location-based 'fast, fun and future' {link} #SXSW #SXSWi
2096    RT @mention Geeking out on YouTube APIs #SXSW @mention Google Teaching Theatre {link}
925     ballroom d: #marissagoogle talking about some cool projects (obv). love the Google Art Project. #sxsw
2987    P.S. @mention and Google throw a b!tchin' party. Shout out to The Spazmatics #sxsw
250     Loved the honesty in Google's Marissa Mayer keynote: we have too many products and need to step up customer service for locations #
2130    RT @mention Google to Launch Major New Social Network Called Circles {link} #sxsw / cc @mention @mention
2489    Leaving Google's Marissa Mayer Keynote. Interesting details on user adoption of location-aware services. #sxsw
2099    RT @mention Get it while it's hot! The latest version of Whrrl is available today for Android, iPhone - and Blackberry! WHRRL FOR B
1473    Great talk on using game mechanics to get user engagement @mention from google rocked it. #sxsw
912     When brand focuses on purpose, not object, they survive &amp;amp; succeed. Google: not search, useful info. Nike: not sneakers, perform
78      Just left #sxsw tradeshow demo of @mention at the Google Theatre. Ok, I get it. I see why all the presenters here are using it.
666     Watching a promo for Google earth engine at 'Techies can save the world, why don't they?'. Harnessing collective power for good. #s
19      Okay, this is really it: yay new @Foursquare for #Android app!!!!11 kthxbai. #sxsw
2976    @mention this time next week Google party at #SXSW!
1084    If you aren't at google you just missed the dance party of a lifetime #SXSW
141     And a few are Android too RT @mention 10 New Mobile Apps I  Ûªll Be Using atåÊ#SXSW  {link}
2322    RT @mention RT @mention :) RT @mention &quot;Google before you tweet&quot; is the new &quot;think before you speak.&quot; - Mark Be
anel at #SXSW.
Name: text, dtype: object
```

SUMMARY of Google positive tweets

```
Android fan
old Galaxy S running Android 2.1

Android app... for ? coming out; FourSquare for Android;
best Android app ... Gowalla 3.0; (sweet; enjoying changes)

google calendar for conference (parties and showcases)
google groups - PartnerHub for conference
meet google pepople - face to the company
Impromptu parties; party w/ Google
Barry Diller at party

Marisa Mayer keynote: location aware apps / services; better CS

open systems
better cloud
checkins feature
google local and search
3-D google maps / google maps navigation
game mechanics - google presenter

Mark Belinsky quote - think before you tweet
```

n [34]:
```
data5 = data2.loc[(data2['brand'] == 'Apple') & (data2['feelings'] == 0)]
print(data5.shape)
data5.head()
```

```
(387, 4)
```

ut[34]:

	text	brand	feelings	text_processed
0	.@wesley83 I have a 3G iPhone. After 3 hrs twe...	Apple	0	@wesley83 i have a 3g iphone after 3 hrs tweet...
3	@sxsw I hope this year's festival isn't as cra...	Apple	0	@sxsw i hope this year's festival isn't as cra...
14	I just noticed DST is coming this weekend. How...	Apple	0	i just noticed dst is coming this weekend how ...
38	attending @mention iPad design headaches #sxsw...	Apple	0	attending @mention ipad design headaches #sxsw...
48	What !?!? @mention #SXSW does not provide iPh...	Apple	0	what @mention #sxsw does not provide iphone ...

n [42]:
```
data5.text[1:25]
```

ut[42]:
```
3      @sxsw I hope this year's festival isn't as crashy as this year's iPhone app. #sxsw
14     I just noticed DST is coming this weekend. How many iPhone users will be an hour late at SXSW come Sunday morning? #SXSW #iPhone
38     attending @mention iPad design headaches #sxsw {link}
48     What !?!? @mention  #SXSW does not provide iPhone chargers?!?  I've changed my mind about going next year!
81     Seriously #sxsw? Did you do any testing on the mobile apps? Constant iPad crashes causing lost schedules, and no sync for WP7.
83     ipad2 and #sxsw...a conflagration of doofusness.  {link}
87     You spent $1,000+ to come to SXSW. \n\nYou've already used iPad 1. \n\nThe wait is a couple city blocks. \n\nWhy? #ipad2 #SXSW {link
90     I'm up to 2 iPad 2s seen in the wild. Both people say it is fast, but the still pics are terrible. #sxsw
104    If iPhone alarms botch the timechange, how many #SXSW'ers freak? Late to flights, missed panels, behind on bloody marys...
106    I meant I also wish I  at #SXSW #dyac stupid iPhone!
113    Found the app kyping my iPhone's geolocation &amp; not releasing when in background. Need a patch, @mention #batterykiller #SXSW
116    Of course Apple built a temp store in Austin. It's Texas. They understand the concept of corralling cattle #SXSW #PickMeUpAniPad2
123      ÛÏ@mention Apple is opening up a temporary store in downtown Austin for #SXSW and the iPad 2 launch&quot; oh YAY more traffic.
132    &quot;The Apple store at the mall on Sunday is 10x as crowded as this. This line is fake. I just need a fucking dongle.&quot; Genius
138    iPad news apps 'so last year' at #SXSW {link}
144    Overheard at #sxsw interactive: &quot;Arg! I hate the iphone! I want my blackberry back&quot; #shocked
145      ÛÏ@mention at #sxsw: &quot;apple comes up with cool technology no one's ever heard of because they don't go to conferences&quot; (
150    overheard at MDW (and I'll second it) &quot;halfway through my iPhone battery already and I haven't even boarded the plane to #sxsw&
158    .@mention Bad Apple: shows up late, Qs the process, poo-poos ideas, leaves early. Can even be &quot;I'm too creative&quot; or busy #
159    Trying to balance the power of power needs on iPhone vs iPad at #sxsw. This 3G iPad sucks it out quick. Might have go airplane mode.
161    My iPhone battery can't keep up with my tweets! Thanks Apple.  #SXSW  #precommerce
165      ÛÏ@mention Best thing I've heard this weekend at #SXSW &quot;I gave my iPad 2 money to #Japan relief. I don't need an iPad 2.&quot
177    iPad 2 is coming out at #SXSW, guess Apple's pretty desperate to give it attention.
181    IPhone is dead. Find me on the secret batphone #sxsw.
Name: text, dtype: object
```

n [75]:
```
data5.text.sample(20)
```

ut[75]:
```
2389     RT @mention The iPad 2 is the also a cartoonishly large digital camera. #SXSW #CStejas {link}
2910     Looking at the line for the pop up #sxsw apple store...I can't think of a single object I want that much.
2476     Fuck. iPhone crapped out, will not charge at all. Says it is charging, but at 5% after all night. What are my options in Austin? #s
2701     Horrible repressed memories of the Apple spinning beach ball coming back at the #progressbar talk. #sxsw
382      I left my pocket guide at the hotel. I don't know how I'm going to cope. What does that say about the usability of iPad/iPhone app?
869      Just launched the pop-up Apple Store at #SXSW. It's our &quot;vintage&quot; store format: Mostly iPods and snarky employees. Ah, th
377      @mention  packing a point by showing iphone fragmentation  #SXSW
1790     I am inventing a &quot;dislike&quot; button for #iPad 2 lines. {link} #SXSW
1826     Decided to go to LA instead of #SXSW, because my AT&amp;T iPhone would be about as useful as a brick in Austin.
1043     Several years too late? I think the trend of social apps is over... @mention  #sxsw #google #circles #conversation @mention
1536     &quot;Apple: the most elegant fascist corporation in America today.&quot; -- Kara Swisher #sxsw #flipboard
1800     Grrrr @mention not muting #sxsw as thought on web or iphone :(
2895     Off to get my badge. Then to find food and drink. Then figure out why my @mention iPhone is NOT roaming at #sxsw. Then unpack. Prio
1085     To my friends at #SXSW who think I abandoned you, in reality I just didn't have any means of communication, my iPhone stopped worki
1654     Please can I hear more people talk about the iPad 2, preferrably with more #sxsw hashtags -  I really have a deficit of this in my
145        ÛÏ@mention at #sxsw: &quot;apple comes up with cool technology no one's ever heard of because they don't go to conferences&quot;
2472     #sxsw iPhone app: control mania! Half of the screen used for buttons and filters, other half for content. #ui-fail
3123     #iPad #news #apps not popular with the #kids. {link} #the_daily is a terrible concept anyway #sxsw
1238     @mention good job @mention #sxsw! went home &amp; watched season 1 of the guild =D. sucks that your tweet abt the iphone hijack is
1418     @mention is about to talk about the mistakes he made building Netflix for the iPhone. #SXSW #netflixiphone
Name: text, dtype: object
```

SUMMARY of Apple negative tweets

poor battery life iPhone? ; iPad 3G sucks power fast ; battery not charging; iPhone stopped working
lack of iphone chargers / charging stations /
poor phone service in Austin with ATT on iPhone ; iPhone not roaming;
general dislike iPhone - want blackberry back
bugs and errors with iPhone - auto-complete; provide geo-location
poor still pics on iPad2; cartonishly large digital camera

crashy iPhone app / apps crashing on iPad / prefer printed version(conference app)
poor time zone / time change correcting on iPhone

iPad2 launch - desperate for attention; overloaded with hype and hashtags;
line is long at Apple store; long lines for iPad2; lines;
temporary store - downtown Austin
poor cust.service at pop-up store (snarky employees)

design headaches preso ; poor design of Aple app (control mania)
ui issues - progress bar;
Netflix for he iphone (app mistakes?)
iPad news app - so last year; not popular
Japan relief - gave money so cant afford iPad2

general company badness (Kara Swisher quote - Apple, the most elegant fascist corporation in America today)
Apple lacks presence at conferences

n [35]:
```
data6 = data2.loc[(data2['brand'] == 'Google') & (data2['feelings'] == 0)]
print(data6.shape)
data6.head()
```

```
(131, 4)
```

ut[35]:

	text	brand	feelings	text_processed
30	@mention - False Alarm: Google Circles Not Co...	Google	0	@mention - false alarm: google circles not co...
157	they took away the lego pit but replaced it wi...	Google	0	they took away the lego pit but replaced it wi...
168	Google vs Bing on #bettersearch. Bing has a sh...	Google	0	google vs bing on #bettersearch bing has a sho...
238	ÛÏ@mention Google to Launch Major New Social ...	Google	0	ûï@mention google to launch major new social ...
269	google is interested in location based tech fo...	Google	0	google is interested in location based tech fo...

n [40]:
```
pd.set_option('display.max_colwidth', -1)
```

n [41]:
```
data6.text[1:25]
```

ut[41]:
```
157    they took away the lego pit but replaced it with a recharging station ;) #sxsw and i might check prices for an iphone - crap samsung
168    Google vs Bing on #bettersearch. Bing has a shot at success w/ structured search.  Potentially higher margin CPA model vs #Google. #
238      ÛÏ@mention Google to Launch Major New Social Network Called Circles, Possibly Today {link} #sxsw  Û   \nIt'll never beat myspace.
269    google is interested in location based tech for indoor venues - businesses, convention centers etc.  Tech needs to improve first.  #
302    more that just location, PixieEngine! RT @mention Google says the future is location, location, location: {link} #SXSW #CNN
307    Google to Launch Major New Social Network Called Circles (Updated) {link} *Not launched at #SXSW, but soon. Should I care?
321    Google to launch product! Wait, no launch, but product exists.  Wait, product does not exist! #sxsw {link}
347    compiling my #sxsw list in one google doc is taking a lot longer than i thought... so many parties. so many good musicians.
390    .@mention Problem with Google Living Stories was the process of creating content didn't change - was just an interface. #hacknews #S
399    L.A.M.E.  RT @mention &quot;...by the law of averages, better than Buzz&quot; RT @mention &quot;Google Circles will be _____ &quot;
411    Q: Why do social sites like Delicious often have better results than Google or Bing? #qagb #sxsw
466    So we get to see google fail at social on another day RT @mention Okay, no Google Circles debuting at #sxsw today
470    #futuremf Trajan: Google has destroyed the &lt;title&gt; tag - websites SEO them. Open Graph Protocol added a clean title tag instea
480    Dense una vuelta por #socialfuel #sxsw para ver la gran diferencia..RT @mention &quot;The revolution will be clumsily translated by
494    Just when you thought &quot;social&quot; couldn't get more overblown at #sxsw, Google may be announcing &quot;Circles&quot; today: {
522    So true!!! RT @mention 'Google lost its way by caring too much for the business vs. the users' - @mention #psych #sxsw
589    @mention &quot;I've worked at Google for over 11 years, so I've seen a lot of evil.&quot; #qagb #sxsw
636    #SXSW 2011: The #Google and #Bing smackdown in all its bloody banality {link}
714    Why does all the #Android meetups here in #Austin are when I'm at work. Well at least there is the PS meetup #sxsw
733    @mention Android needs a way to group apps like you can now do with iPad/iPod. #SXSW #hhrs
739    @mention another google social failure? #sxsw
780    Guy just asked Google's Mayer: it can take a year to remove deadly routes from Google Maps, such as through Death Valley #SXSW
831    Marissa Mayer: Google maps should have better customer service, quicker responses. #sxsw #FH
855    The walk by Lady Bird Lake was lovely, but Google Maps travel times are not to be trusted. #SXSW
Name: text, dtype: object
```

n [76]:
```
data6.text.sample(20)
```

ut[76]:
```
1563    Google will eat itself #rhizome #sxswk #sxsw  @mention Hilton {link}
831     Marissa Mayer: Google maps should have better customer service, quicker responses. #sxsw #FH
1101    Ha! RT @mention Google guy at #sxsw talk is explaining how he made realistic Twitter bots as an experiment. Gee, thanks for doing t
1239    My #sxsw Google calendar is getting a little out of control
347     compiling my #sxsw list in one google doc is taking a lot longer than i thought... so many parties. so many good musicians.
168     Google vs Bing on #bettersearch. Bing has a shot at success w/ structured search.  Potentially higher margin CPA model vs #Google.
1986    RT @mention And it will suck. RT @mention RT @mention Google will preview major new social service, Circles, at #sxsw today {link}
2525    Looking forward to the day when @mention and @mention release native Android 3.0 tablet-optimized clients! Google Latitude sucks! #
466     So we get to see google fail at social on another day RT @mention Okay, no Google Circles debuting at #sxsw today
238       ÛÏ@mention Google to Launch Major New Social Network Called Circles, Possibly Today {link} #sxsw  Û   \nIt'll never beat myspace.
1635    Nope, seems no Google Circles launch today: {link} #sxsw
1229    &quot;Google to Launch Major New Social Network.&quot; really dont need another social network...{link} #sxsw
2488    Damn it Google! Your glow-in-the-dark cup leaked glow-in-the-dark goo in my camera bag! #SXSW
470     #futuremf Trajan: Google has destroyed the &lt;title&gt; tag - websites SEO them. Open Graph Protocol added a clean title tag inste
302     more that just location, PixieEngine! RT @mention Google says the future is location, location, location: {link} #SXSW #CNN
2116    RT @mention Google hotpot- rate restaurants and get personalized recos on where to eat. Um, think foursquare, yelp, etc have this c
W
2166    RT @mention Hm? Do we need another 1? RT @mention Google to Launch Major New Social Network Called Circles, Possibly Today {link} #
2123    RT @mention Google Latina and see what you find? Porn...this is the first impression that people get about us? #latism #sxsw #sxswL
h  Û
1851    Google product showcases never feel that cool. No price tag, brand equity, wow factor attached. #sxsw marissa mayer
1392    Hey @mention got invited to a new group at #SXSW and your Android app keeps crashing when I try to join! WTF? #sxswfail
Name: text, dtype: object
```

SUMMARY of Google negative tweets

samsung android phone dislike

google maps - takes long to fix; untrusted travel times; needs better cust. service;
Marisa Mayer talk (long time to fix google maps issues - routing people to the center of Death Valley)
need location based tech for indoor venues

search compared to Bing (#bettersearch); Bing has better structured search; better cpa model
search results BAD - poor results when searhc on Latina (get porn)
goolge SEO distroyed utility of title tag

new social netwrok - google circles (lame; not as good as mySpace; fail; overblown; it will suck;
others do restaurant recoos better - Google hotspot

google docs slow
google latitude sucks...
google Living Stories - did not change content creation process
google translate is clumsy

need way to group apps - like iPad

```
need Android3 tablet optimized clients
Android app crashes - for new group?

too business focused instead of users
is evil
focus on wring things (twitter bots experiment)

free cup leaked goo in camera bag
```

Observations on the tweets
So we see that we can glean more meaning from the word based modeling by reading our data (aka the tweets). Perhaps reading is to NLP as ED visualizations are to numerical datasets / modeling. Above I have summarized the main themes for the 4 groups of tweets.

In [ ]:

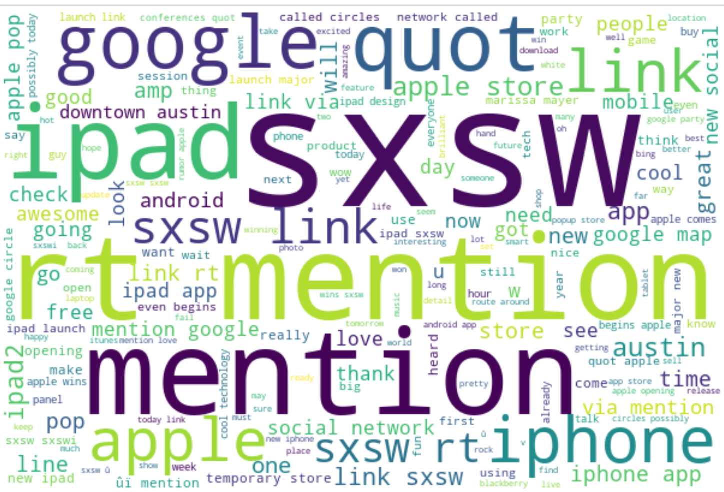
0. Full Data Set into LDA
Initially, went through the LDA process with the full tweet set of 3182 to make sure it worked. As expected it was not too insightful. Also initially I had not removed the additional unique stopwords (32) to get rid of the very prevelant words.

In [5]:
```python
# TEXT Pre-processing
# Load the regular expression library
import re
# Remove punctuation
data2['text_processed'] = \
data2['text'].map(lambda x: re.sub('[,\.!?]', '', x))
# Convert the titles to lowercase
data2['text_processed'] = \
data2['text_processed'].map(lambda x: x.lower())
# Print out the first rows of papers
data2['text_processed'].head()
```

Out[5]:
```
0    @wesley83 i have a 3g iphone after 3 hrs tweet...
1    @jessedee know about @fludapp  awesome ipad/ip...
2    @swonderlin can not wait for #ipad 2 also they...
3    @sxsw i hope this year's festival isn't as cra...
4    @sxtxstate great stuff on fri #sxsw: marissa m...
Name: text_processed, dtype: object
```

[104]:
```python
# EDA word viz... for whole dataset.
# Import the wordcloud library
from wordcloud import WordCloud
# Join the different processed titles together.
long_string = ','.join(list(data2['text_processed'].values))
# Create a WordCloud object
wordcloud = WordCloud(background_color="white", max_words=200, height=400, width=600, contour_width=3, contour_color='steelblue')
# Generate a word cloud
wordcloud.generate(long_string)
# Visualize the word cloud
wordcloud.to_image()
```

t[104]:



In [8]:
```
# Prepare the data for LDA analysis
import gensim
from gensim.utils import simple_preprocess
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords

stop_words = stopwords.words('english')
stop_words.extend(['sxsw', 'quot', 'mention', 'link', 'rt', 'amp', 'http' ,'sxswrt'])

def sent_to_words(sentences):
    for sentence in sentences:
        # deacc=True removes punctuations
        yield(gensim.utils.simple_preprocess(str(sentence), deacc=True))
def remove_stopwords(texts):
    return [[word for word in simple_preprocess(str(doc))
             if word not in stop_words] for doc in texts]
data = data2.text_processed.values.tolist()
data_words = list(sent_to_words(data))
# remove stop words
data_words = remove_stopwords(data_words)
print(data_words[:1][0][:30])
```

```
[nltk_data] Downloading package stopwords to /Users/markp/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

```
['wesley', 'iphone', 'hrs', 'tweeting', 'rise_austin', 'dead', 'need', 'upgrade', 'plugin', 'stations']
```

In [9]:
```
# Create dictionary and corpus
import gensim.corpora as corpora
# Create Dictionary
id2word = corpora.Dictionary(data_words)
# Create Corpus
texts = data_words
# Term Document Frequency
corpus = [id2word.doc2bow(text) for text in texts]
# View
print(corpus[:1][0][:30])
```

```
[(0, 1), (1, 1), (2, 1), (3, 1), (4, 1), (5, 1), (6, 1), (7, 1), (8, 1), (9, 1)]
```

n [15]:
```
# Train the model
from pprint import pprint
# number of topics
num_topics = 4
# Build LDA model
lda_model = gensim.models.LdaMulticore(corpus=corpus,
                                       id2word=id2word,
                                       num_topics=num_topics)
# Print the Keyword in the 10 topics
pprint(lda_model.print_topics())
doc_lda = lda_model[corpus]
```

```
[(0,
  '0.042*"ipad" + 0.023*"google" + 0.021*"iphone" + 0.021*"apple" + '
  '0.020*"app" + 0.010*"store" + 0.009*"new" + 0.006*"people" + '
  '0.005*"android" + 0.005*"day"'),
 (1,
  '0.038*"ipad" + 0.025*"apple" + 0.023*"iphone" + 0.022*"google" + '
  '0.018*"new" + 0.016*"store" + 0.013*"app" + 0.007*"android" + 0.007*"get" + '
  '0.006*"free"'),
 (2,
  '0.020*"ipad" + 0.016*"apple" + 0.016*"google" + 0.010*"iphone" + '
  '0.007*"store" + 0.004*"via" + 0.004*"great" + 0.004*"win" + 0.004*"coming" '
  '+ 0.003*"free"'),
 (3,
  '0.040*"apple" + 0.038*"ipad" + 0.022*"google" + 0.022*"store" + '
  '0.021*"austin" + 0.013*"iphone" + 0.011*"pop" + 0.010*"launch" + '
  '0.007*"opening" + 0.007*"downtown"')]
```

n [17]:
```
# Code from Yish
print(lda_model.print_topics())
```

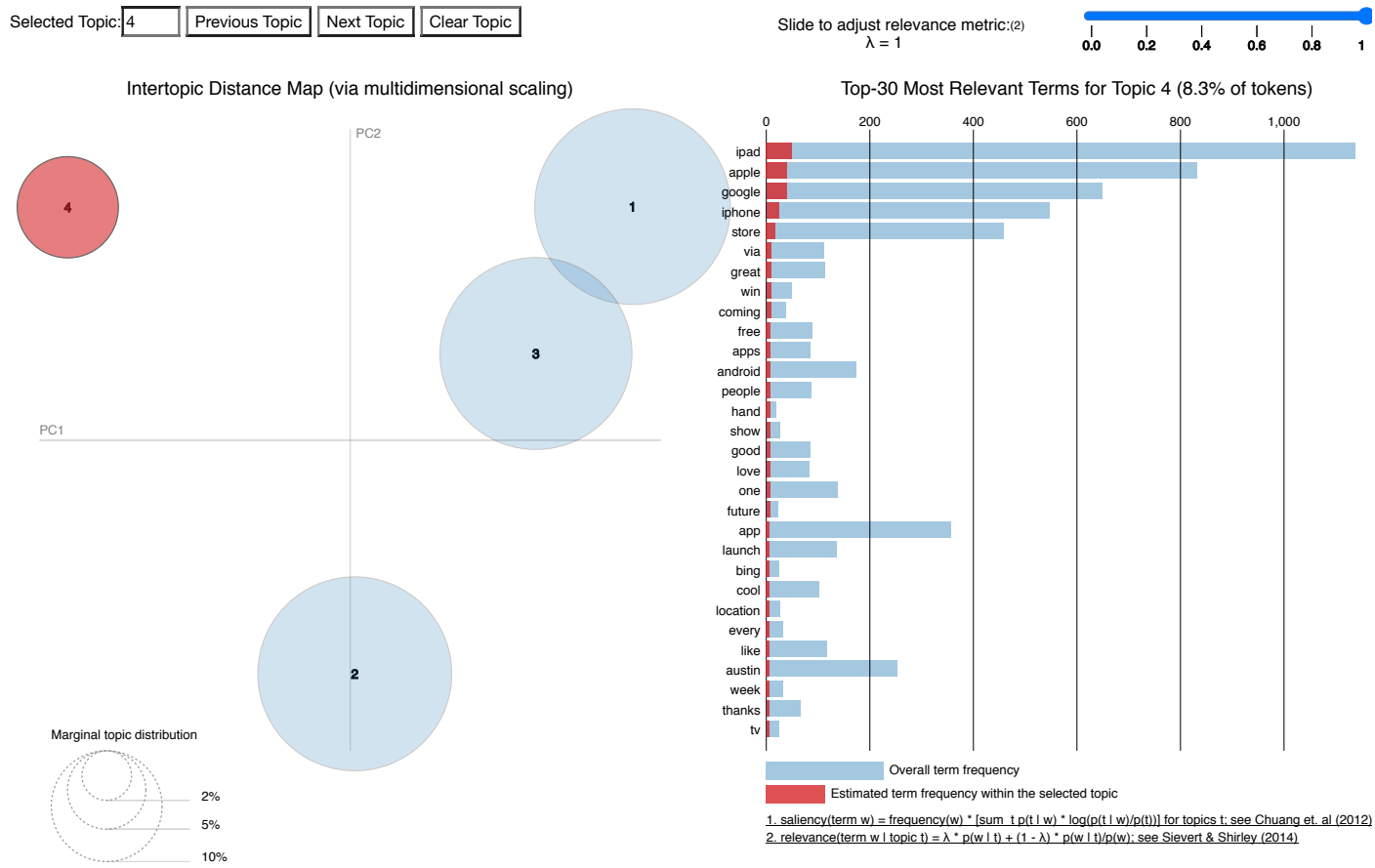
```
[(0, '0.042*"ipad" + 0.023*"google" + 0.021*"iphone" + 0.021*"apple" + 0.020*"app" + 0.010*"store" + 0.009*"new" + 0.006*"people" + 0.005*"
y"'), (1, '0.038*"ipad" + 0.025*"apple" + 0.023*"iphone" + 0.022*"google" + 0.018*"new" + 0.016*"store" + 0.013*"app" + 0.007*"android" + 0
*"free"'), (2, '0.020*"ipad" + 0.016*"apple" + 0.016*"google" + 0.010*"iphone" + 0.007*"store" + 0.004*"via" + 0.004*"great" + 0.004*"win"
0.003*"free"'), (3, '0.040*"apple" + 0.038*"ipad" + 0.022*"google" + 0.022*"store" + 0.021*"austin" + 0.013*"iphone" + 0.011*"pop" + 0.010*
ening" + 0.007*"downtown"')]
```

n [18]:
```
# Code to generate the interactive vizualization
import pyLDAvis.gensim
pyLDAvis.enable_notebook()
```

n [19]:
```
vis = pyLDAvis.gensim.prepare(lda_model, corpus, id2word)
vis
```

```
/Users/markp/opt/anaconda3/envs/learn-env/lib/python3.6/site-packages/pyLDAvis/_prepare.py:257: FutureWarning: Sorting because non-concaten
igned. A future version
of pandas will change to not sort by default.

To accept the future behavior, pass 'sort=False'.

To retain the current behavior and silence the warning, pass 'sort=True'.

  return pd.concat([default_term_info] + list(topic_dfs))
```

ut[19]:



Observations on the initial LDA

Note that this initial Topic Model was with the entire dataset. Initially I tried a grouping of 10. The top 6 salient terms were all compan produt names. And it was odd that all 10 of the topics contained both Apple and Google names... there was no distinction. Also from the rem words, it was difficult to dettermine which were negative terms... in this context. Need to keep in mind that overall only 16% of the tweet negative, so this is reflected. Also tried (and displayed above 4 topics) there appears to be spacial distance in Topics 2 and 4, but diffu pick out what is going on just by the words.

In [ ]:

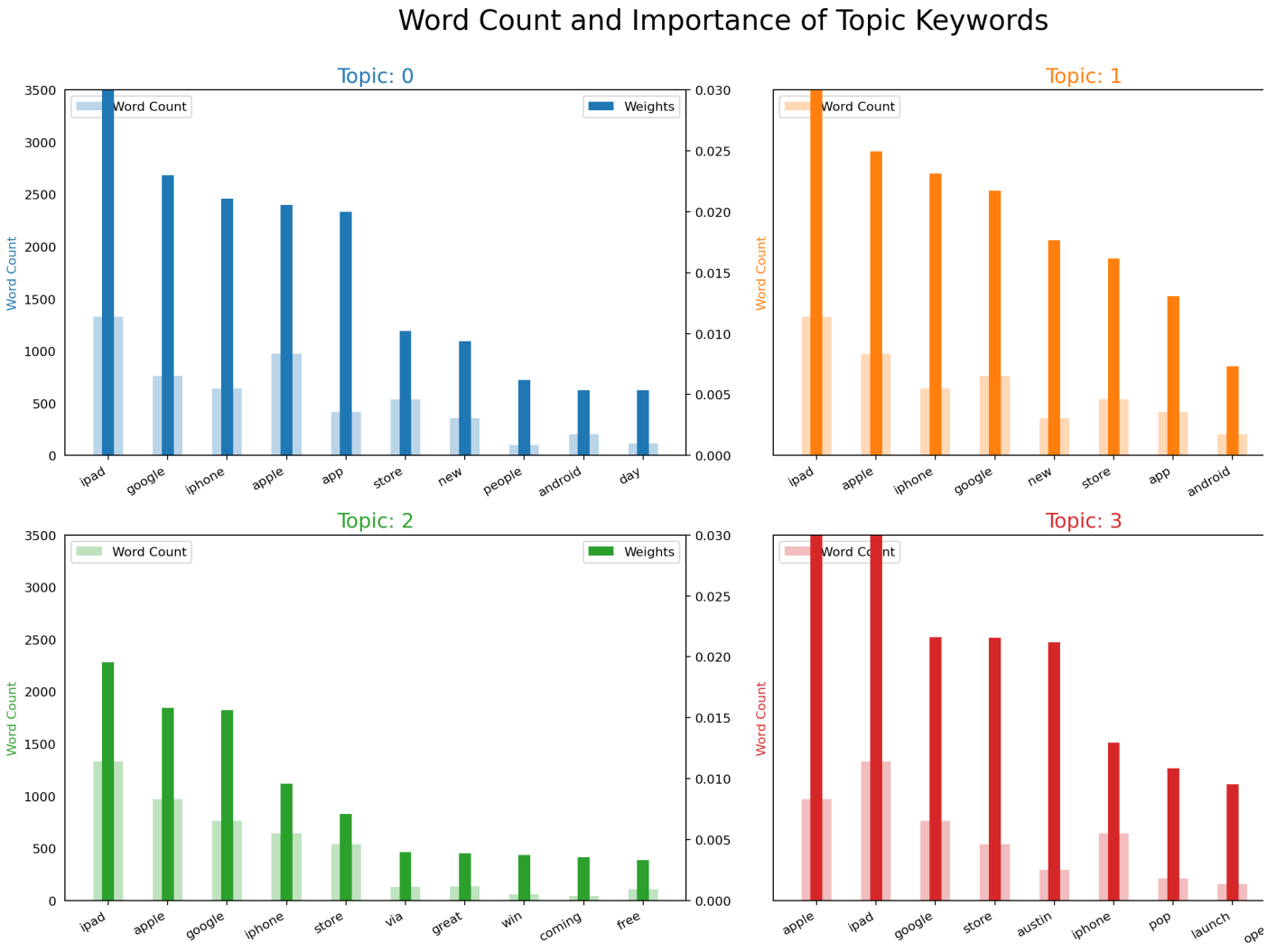
Will try some other visualizations - still whole dataset

These are from the S.V. article on machinelearningplus.com

n [22]:
```
import matplotlib.colors as mcolors
# cols = [color for name, color in mcolors.TABLEAU_COLORS.items()]  # more colors: 'mcolors.XKCD_COLORS'
```

n [26]:
```
# Word count of Topic Keywords (can compare weights versus word counts).

from collections import Counter
topics = lda_model.show_topics(formatted=False)
data_flat = [w for w_list in data_words for w in w_list]
counter = Counter(data_flat)

out = []
for i, topic in topics:
    for word, weight in topic:
        out.append([word, i , weight, counter[word]])

df = pd.DataFrame(out, columns=['word', 'topic_id', 'importance', 'word_count'])

# Plot Word Count and Weights of Topic Keywords
fig, axes = plt.subplots(2, 2, figsize=(16,10), sharey=True, dpi=160)
cols = [color for name, color in mcolors.TABLEAU_COLORS.items()]
for i, ax in enumerate(axes.flatten()):
    ax.bar(x='word', height="word_count", data=df.loc[df.topic_id==i, :], color=cols[i], width=0.5, alpha=0.3, label='Word Count')
    ax_twin = ax.twinx()
    ax_twin.bar(x='word', height="importance", data=df.loc[df.topic_id==i, :], color=cols[i], width=0.2, label='Weights')
    ax.set_ylabel('Word Count', color=cols[i])
    ax_twin.set_ylim(0, 0.030); ax.set_ylim(0, 3500)
    ax.set_title('Topic: ' + str(i), color=cols[i], fontsize=16)
    ax.tick_params(axis='y', left=False)
    ax.set_xticklabels(df.loc[df.topic_id==i, 'word'], rotation=30, horizontalalignment= 'right')
    ax.legend(loc='upper left'); ax_twin.legend(loc='upper right')

fig.tight_layout(w_pad=2)
fig.suptitle('Word Count and Importance of Topic Keywords', fontsize=22, y=1.05)
plt.show()
```

```
/Users/markp/opt/anaconda3/envs/learn-env/lib/python3.6/site-packages/ipykernel_launcher.py:26: UserWarning: FixedFormatter should only be
ixedLocator
```



In [ ]:
```
# Sentence chart colored by topic - ran out of time to try this...
```

In [ ]:
```

```

1. Try seperate LDA for negative tweets (Apple)

This is data5 and has 387 tweets.

[106]:
```
# data5.head()
```

n [46]:
```
# TEXT Pre-processing
# Load the regular expression library
import re
# Remove punctuation
data5['text_processed'] = \
data5['text'].map(lambda x: re.sub('[,\.!?]', '', x))
# Convert the titles to lowercase
data5['text_processed'] = \
data5['text_processed'].map(lambda x: x.lower())
# Print out the first rows
data5['text_processed'].head()
```

```
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<ipython-input-46-16dd6e38873b>:5: DeprecationWarning: invalid escape sequence \.
  data5['text_processed'] = data5['text'].map(lambda x: re.sub('[,\.!?]', '', x))
/Users/markp/opt/anaconda3/envs/learn-env/lib/python3.6/site-packages/ipykernel_launcher.py:5: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (
a.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  """
/Users/markp/opt/anaconda3/envs/learn-env/lib/python3.6/site-packages/ipykernel_launcher.py:7: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (
a.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  import sys
```

ut[46]:
```
0     @wesley83 i have a 3g iphone after 3 hrs tweeting at #rise_austin it was dead  i need to upgrade plugin stations at #sxsw
3     @sxsw i hope this year's festival isn't as crashy as this year's iphone app #sxsw
14    i just noticed dst is coming this weekend how many iphone users will be an hour late at sxsw come sunday morning #sxsw #iphone
38    attending @mention ipad design headaches #sxsw {link}
48    what  @mention  #sxsw does not provide iphone chargers  i've changed my mind about going next year
Name: text_processed, dtype: object
```

n [97]:
```
stop_words = stopwords.words('english')
stop_words.extend(['sxsw', 'sxswi', 'store', 'quot', 'mention', 'link', 'rt', 'amp', 'http', 'sxswrt', 'google', 'googles', 'app', 'apps',
```

n [98]:
```
# EDA word viz... for data5...Apple negative.
# Import the wordcloud library
from wordcloud import WordCloud
# Join the different processed titles together.
long_string = ','.join(list(data5['text_processed'].values))
# Create a WordCloud object
wordcloud = WordCloud(background_color="white", stopwords=stop_words, max_words=200, height=400, width=600, contour_width=3, contour_color=
# Generate a word cloud
wordcloud.generate(long_string)
# Visualize the word cloud
wordcloud.to_image()
```

ut[98]:

[107]:
```
# Prepare the data for LDA analysis
import gensim
from gensim.utils import simple_preprocess
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords

stop_words = stopwords.words('english')
stop_words.extend(['sxsw', 'sxswi', 'store', 'quot', 'mention', 'link', 'rt', 'amp', 'http', 'sxswrt', 'google', 'googles', 'app', 'apps',
def sent_to_words(sentences):
    for sentence in sentences:
        # deacc=True removes punctuations
        yield(gensim.utils.simple_preprocess(str(sentence), deacc=True))
def remove_stopwords(texts):
    return [[word for word in simple_preprocess(str(doc))
             if word not in stop_words] for doc in texts]
data = data5.text_processed.values.tolist()
data_words = list(sent_to_words(data))
# remove stop words
data_words = remove_stopwords(data_words)
print(data_words[:1][0][:30])
```

```
['wesley', 'hrs', 'tweeting', 'rise_austin', 'dead', 'need', 'upgrade', 'plugin', 'stations']

[nltk_data] Downloading package stopwords to /Users/markp/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

[108]:
```
# Create dictionary and corpus
import gensim.corpora as corpora
# Create Dictionary
id2word = corpora.Dictionary(data_words)
# Create Corpus
texts = data_words
# Term Document Frequency
corpus = [id2word.doc2bow(text) for text in texts]
# View
print(corpus[:1][0][:30])
```

```
[(0, 1), (1, 1), (2, 1), (3, 1), (4, 1), (5, 1), (6, 1), (7, 1), (8, 1)]
```

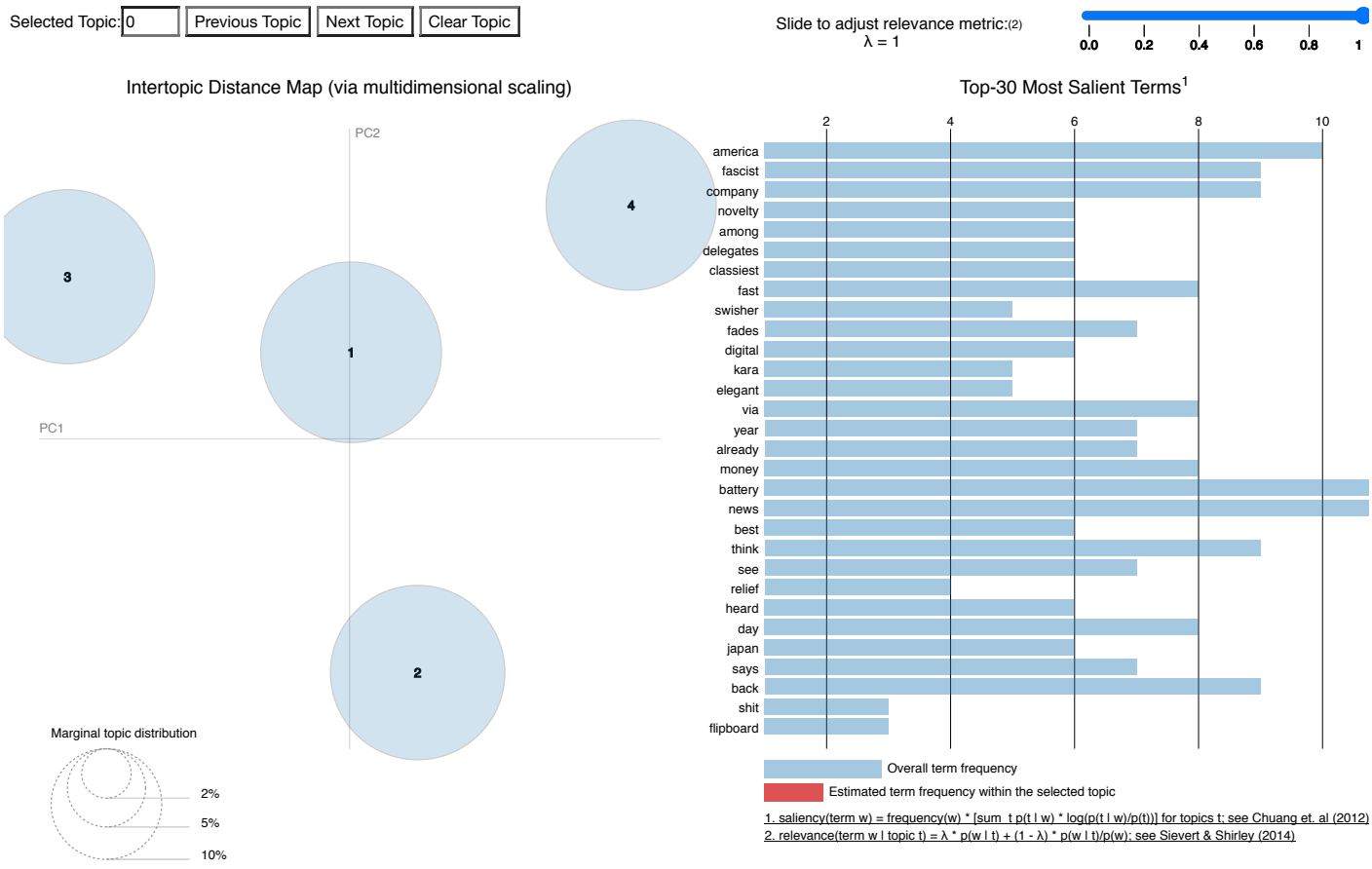
[109]:
```
# Train the model
from pprint import pprint
# number of topics
num_topics = 4
# Build LDA model
lda_model = gensim.models.LdaMulticore(corpus=corpus,
                                       id2word=id2word,
                                       num_topics=num_topics)
# Print the Keyword in the 4 topics
pprint(lda_model.print_topics())
doc_lda = lda_model[corpus]
```

```
[(0,
  '0.011*"like" + 0.008*"money" + 0.008*"year" + 0.008*"design" + 0.007*"need" '
  '+ 0.006*"best" + 0.006*"japan" + 0.006*"says" + 0.006*"people" + '
  '0.006*"back"'),
 (1,
  '0.014*"america" + 0.013*"fascist" + 0.013*"company" + 0.009*"people" + '
  '0.008*"classiest" + 0.007*"swisher" + 0.007*"get" + 0.006*"kara" + '
  '0.006*"elegant" + 0.006*"see"'),
 (2,
  '0.008*"like" + 0.008*"think" + 0.008*"battery" + 0.007*"already" + '
  '0.006*"pop" + 0.006*"design" + 0.005*"mobile" + 0.005*"line" + 0.005*"yet" '
  '+ 0.005*"many"'),
 (3,
  '0.010*"news" + 0.009*"fast" + 0.008*"fades" + 0.008*"novelty" + '
  '0.008*"delegates" + 0.008*"among" + 0.007*"via" + 0.007*"digital" + '
  '0.006*"fail" + 0.006*"day"')]
```

[110]:
```
# Visualize the resulting clusters
import pyLDAvis.gensim
pyLDAvis.enable_notebook()

vis = pyLDAvis.gensim.prepare(lda_model, corpus, id2word)
vis
```

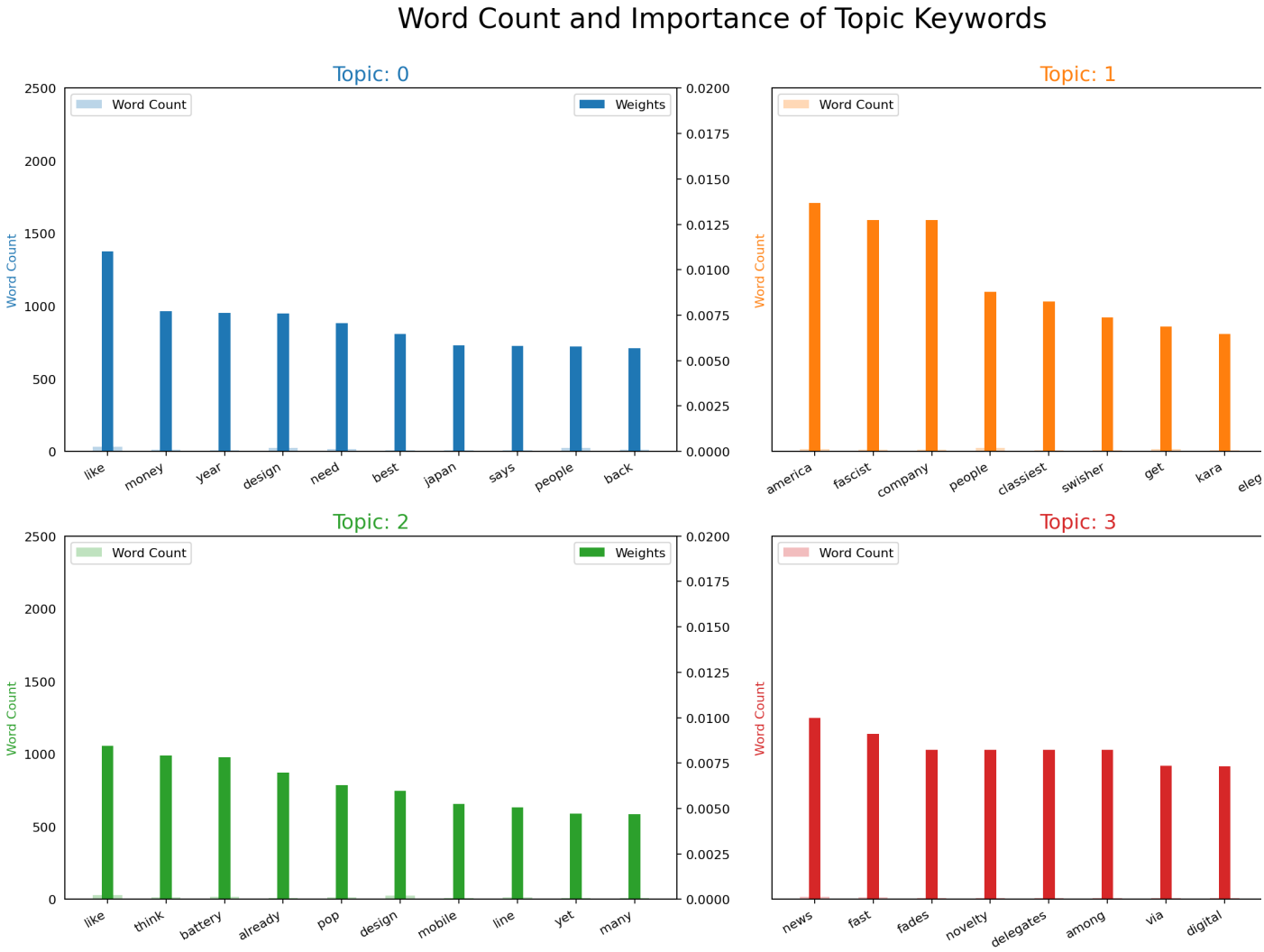
```
/Users/markp/opt/anaconda3/envs/learn-env/lib/python3.6/site-packages/pyLDAvis/_prepare.py:257: FutureWarning: Sorting because non-concaten
igned. A future version
of pandas will change to not sort by default.

To accept the future behavior, pass 'sort=False'.

To retain the current behavior and silence the warning, pass 'sort=True'.

  return pd.concat([default_term_info] + list(topic_dfs))
```

t[110]:



[111]:
```
# Word count of Topic Keywords (can compare weights versus word counts).

from collections import Counter
topics = lda_model.show_topics(formatted=False)
data_flat = [w for w_list in data_words for w in w_list]
counter = Counter(data_flat)

out = []
for i, topic in topics:
    for word, weight in topic:
        out.append([word, i , weight, counter[word]])

df = pd.DataFrame(out, columns=['word', 'topic_id', 'importance', 'word_count'])

# Plot Word Count and Weights of Topic Keywords
fig, axes = plt.subplots(2, 2, figsize=(16,10), sharey=True, dpi=160)
cols = [color for name, color in mcolors.TABLEAU_COLORS.items()]
for i, ax in enumerate(axes.flatten()):
    ax.bar(x='word', height="word_count", data=df.loc[df.topic_id==i, :], color=cols[i], width=0.5, alpha=0.3, label='Word Count')
    ax_twin = ax.twinx()
    ax_twin.bar(x='word', height="importance", data=df.loc[df.topic_id==i, :], color=cols[i], width=0.2, label='Weights')
    ax.set_ylabel('Word Count', color=cols[i])
    ax_twin.set_ylim(0, 0.020); ax.set_ylim(0, 2500)
    ax.set_title('Topic: ' + str(i), color=cols[i], fontsize=16)
    ax.tick_params(axis='y', left=False)
    ax.set_xticklabels(df.loc[df.topic_id==i, 'word'], rotation=30, horizontalalignment= 'right')
    ax.legend(loc='upper left'); ax_twin.legend(loc='upper right')

fig.tight_layout(w_pad=2)
fig.suptitle('Word Count and Importance of Topic Keywords', fontsize=22, y=1.05)
plt.show()
```

```
/Users/markp/opt/anaconda3/envs/learn-env/lib/python3.6/site-packages/ipykernel_launcher.py:26: UserWarning: FixedFormatter should only be
ixedLocator
```



2. Try seperate LDA for negative tweets (Google)

This is data6 and has 131 tweets.

[112]:
```
data6.shape
```

t[112]:
```
(131, 4)
```

[114]:
```
# data6.head()
```

[115]:
```
# TEXT Pre-processing
# Load the regular expression library
import re
# Remove punctuation
data6['text_processed'] = \
data6['text'].map(lambda x: re.sub('[,\.!?]', '', x))
# Convert the titles to lowercase
data6['text_processed'] = \
data6['text_processed'].map(lambda x: x.lower())
# Print out the first rows of papers
data6['text_processed'].head()
```

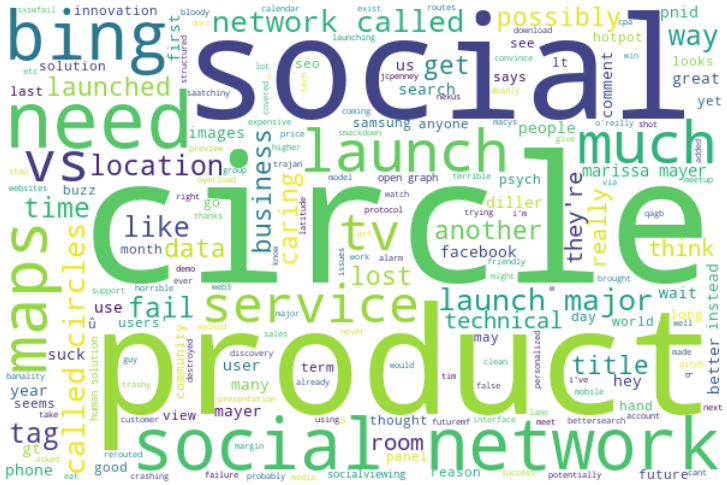
```
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<ipython-input-115-a69b31ada2ba>:5: DeprecationWarning: invalid escape sequence \.
  data6['text_processed'] = data6['text'].map(lambda x: re.sub('[,\.!?]', '', x))
/Users/markp/opt/anaconda3/envs/learn-env/lib/python3.6/site-packages/ipykernel_launcher.py:5: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (
a.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  """
/Users/markp/opt/anaconda3/envs/learn-env/lib/python3.6/site-packages/ipykernel_launcher.py:7: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (
a.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  import sys
```

t[115]:
```
30     @mention  - false alarm: google circles not coming now  ûòand probably not ever - {link} #google #circles #social #sxsw
157    they took away the lego pit but replaced it with a recharging station ;) #sxsw and i might check prices for an iphone - crap samsung
168    google vs bing on #bettersearch bing has a shot at success w/ structured search  potentially higher margin cpa model vs #google #sxs
238      ûï@mention google to launch major new social network called circles possibly today {link} #sxsw  û   \nit'll never beat myspace
269    google is interested in location based tech for indoor venues - businesses convention centers etc  tech needs to improve first  #sxs
Name: text_processed, dtype: object
```

[116]:
```
stop_words = stopwords.words('english')
stop_words.extend(['sxsw', 'sxswi', 'store', 'quot', 'mention', 'link', 'rt', 'amp', 'http', 'sxswrt', 'google', 'googles', 'app', 'apps',
```

[117]:
```
# EDA word viz... for whole dataset.
# Import the wordcloud library
from wordcloud import WordCloud
# Join the different processed titles together.
long_string = ','.join(list(data6['text_processed'].values))
# Create a WordCloud object
wordcloud = WordCloud(background_color="white", stopwords=stop_words, max_words=200, height=400, width=600, contour_width=3, contour_color=
# Generate a word cloud
wordcloud.generate(long_string)
# Visualize the word cloud
wordcloud.to_image()
```

t[117]:



[118]:
```
# Prepare the data for LDA analysis
import gensim
from gensim.utils import simple_preprocess
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords

stop_words = stopwords.words('english')
stop_words.extend(['sxsw', 'sxswi', 'store', 'quot', 'mention', 'link', 'rt', 'amp', 'http', 'sxswrt', 'google', 'googles', 'app', 'apps',

def sent_to_words(sentences):
    for sentence in sentences:
        # deacc=True removes punctuations
        yield(gensim.utils.simple_preprocess(str(sentence), deacc=True))
def remove_stopwords(texts):
    return [[word for word in simple_preprocess(str(doc))
             if word not in stop_words] for doc in texts]
data = data6.text_processed.values.tolist()
data_words = list(sent_to_words(data))
# remove stop words
data_words = remove_stopwords(data_words)
print(data_words[:1][0][:30])
```

```
['false', 'alarm', 'circles', 'coming', 'uoand', 'probably', 'ever', 'circles', 'social']

[nltk_data] Downloading package stopwords to /Users/markp/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

[119]:
```
# Create dictionary and corpus
import gensim.corpora as corpora
# Create Dictionary
id2word = corpora.Dictionary(data_words)
# Create Corpus
texts = data_words
# Term Document Frequency
corpus = [id2word.doc2bow(text) for text in texts]
# View
print(corpus[:1][0][:30])
```

```
[(0, 1), (1, 2), (2, 1), (3, 1), (4, 1), (5, 1), (6, 1), (7, 1)]
```

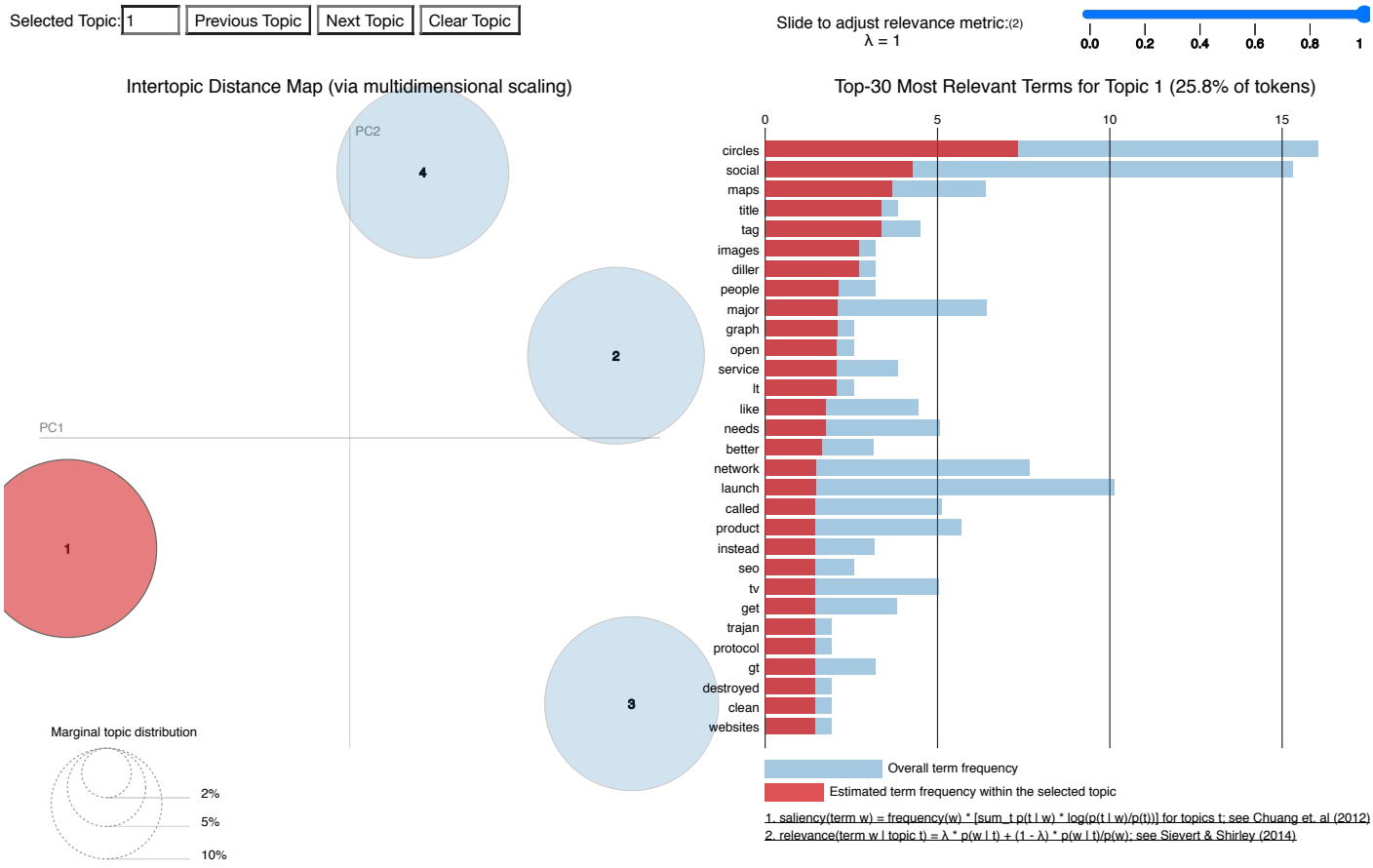
[120]:
```
# Train the model
from pprint import pprint
# number of topics
num_topics = 4
# Build LDA model
lda_model = gensim.models.LdaMulticore(corpus=corpus,
                                       id2word=id2word,
                                       num_topics=num_topics)
# Print the Keyword in the 4 topics
pprint(lda_model.print_topics())
doc_lda = lda_model[corpus]
```

```
[(0,
  '0.025*"circles" + 0.015*"social" + 0.013*"maps" + 0.011*"title" + '
  '0.011*"tag" + 0.009*"images" + 0.009*"diller" + 0.007*"people" + '
  '0.007*"major" + 0.007*"graph"'),
 (1,
  '0.011*"location" + 0.010*"products" + 0.009*"launch" + 0.009*"social" + '
  '0.008*"tv" + 0.008*"mayer" + 0.007*"much" + 0.007*"product" + 0.006*"needs" '
  '+ 0.006*"hey"'),
 (2,
  '0.021*"social" + 0.019*"launch" + 0.019*"network" + 0.018*"circles" + '
  '0.014*"major" + 0.012*"bing" + 0.012*"possibly" + 0.012*"called" + '
  '0.010*"data" + 0.008*"vs"'),
 (3,
  '0.011*"users" + 0.011*"much" + 0.010*"social" + 0.010*"circles" + '
  '0.009*"business" + 0.009*"lost" + 0.009*"caring" + 0.009*"vs" + 0.009*"way" '
  '+ 0.008*"technical"')]
```

[121]:
```
# Visualize the resulting clusters
pyLDAvis.enable_notebook()

vis = pyLDAvis.gensim.prepare(lda_model, corpus, id2word)
vis
```

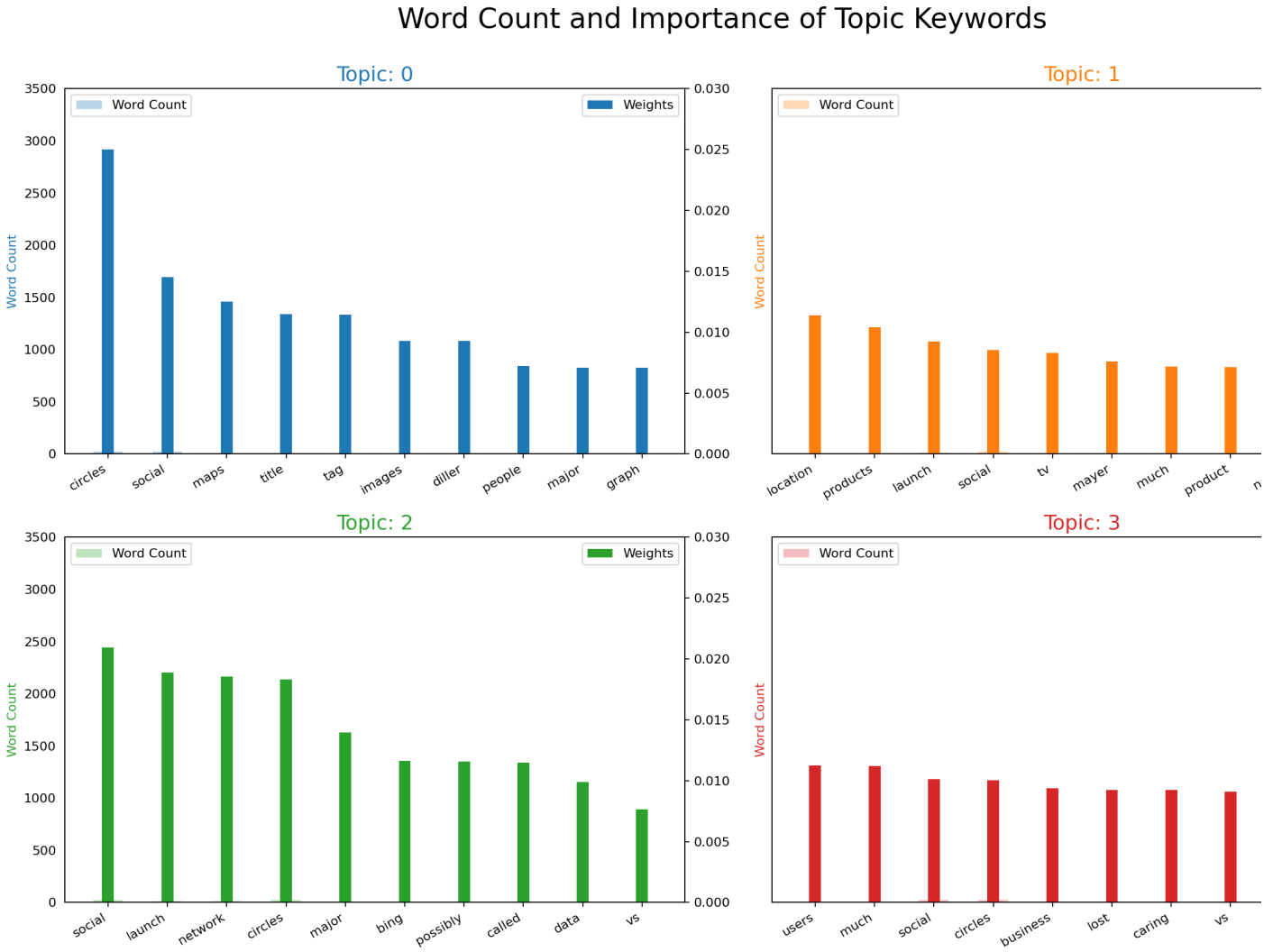
```
/Users/markp/opt/anaconda3/envs/learn-env/lib/python3.6/site-packages/pyLDAvis/_prepare.py:257: FutureWarning: Sorting because non-concaten
igned. A future version
of pandas will change to not sort by default.

To accept the future behavior, pass 'sort=False'.

To retain the current behavior and silence the warning, pass 'sort=True'.

  return pd.concat([default_term_info] + list(topic_dfs))
```

t[121]:



[122]:
```
# Word count of Topic Keywords (can compare weights versus word counts).

from collections import Counter
topics = lda_model.show_topics(formatted=False)
data_flat = [w for w_list in data_words for w in w_list]
counter = Counter(data_flat)

out = []
for i, topic in topics:
    for word, weight in topic:
        out.append([word, i , weight, counter[word]])

df = pd.DataFrame(out, columns=['word', 'topic_id', 'importance', 'word_count'])

# Plot Word Count and Weights of Topic Keywords
fig, axes = plt.subplots(2, 2, figsize=(16,10), sharey=True, dpi=160)
cols = [color for name, color in mcolors.TABLEAU_COLORS.items()]
for i, ax in enumerate(axes.flatten()):
    ax.bar(x='word', height="word_count", data=df.loc[df.topic_id==i, :], color=cols[i], width=0.5, alpha=0.3, label='Word Count')
    ax_twin = ax.twinx()
    ax_twin.bar(x='word', height="importance", data=df.loc[df.topic_id==i, :], color=cols[i], width=0.2, label='Weights')
    ax.set_ylabel('Word Count', color=cols[i])
    ax_twin.set_ylim(0, 0.030); ax.set_ylim(0, 3500)
    ax.set_title('Topic: ' + str(i), color=cols[i], fontsize=16)
    ax.tick_params(axis='y', left=False)
    ax.set_xticklabels(df.loc[df.topic_id==i, 'word'], rotation=30, horizontalalignment= 'right')
    ax.legend(loc='upper left'); ax_twin.legend(loc='upper right')

fig.tight_layout(w_pad=2)
fig.suptitle('Word Count and Importance of Topic Keywords', fontsize=22, y=1.05)
plt.show()
```

```
/Users/markp/opt/anaconda3/envs/learn-env/lib/python3.6/site-packages/ipykernel_launcher.py:26: UserWarning: FixedFormatter should only be
ixedLocator
```



3. Try seperate LDA for positive tweets (Google)

This is data4 and has 719 tweets.

n [55]:
```
data4.shape
```

ut[55]:
```
(719, 4)
```

n [56]:
```
# TEXT Pre-processing
# Load the regular expression library
import re
# Remove punctuation
data4['text_processed'] = \
data4['text'].map(lambda x: re.sub('[,\.!?]', '', x))
# Convert the titles to lowercase
data4['text_processed'] = \
data4['text_processed'].map(lambda x: x.lower())
# Print out the first rows of papers
data4['text_processed'].head()
```

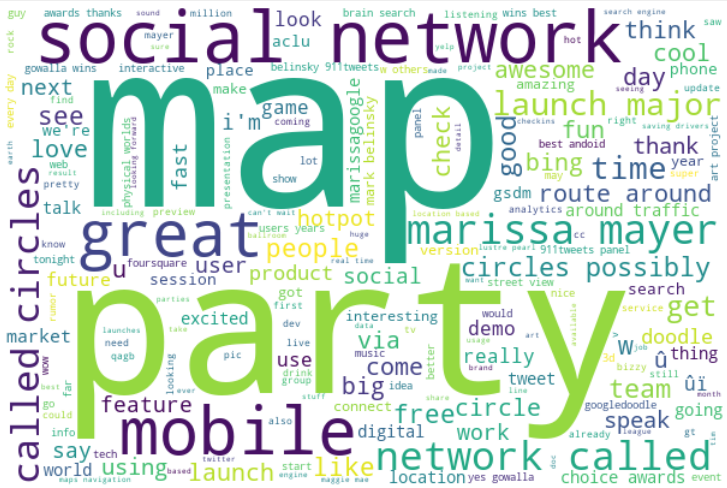
```
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<ipython-input-56-fed53724e348>:5: DeprecationWarning: invalid escape sequence \.
  data4['text_processed'] = data4['text'].map(lambda x: re.sub('[,\.!?]', '', x))
/Users/markp/opt/anaconda3/envs/learn-env/lib/python3.6/site-packages/ipykernel_launcher.py:5: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (
a.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  """
/Users/markp/opt/anaconda3/envs/learn-env/lib/python3.6/site-packages/ipykernel_launcher.py:7: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (
a.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  import sys
```

ut[56]:
```
4     @sxtxstate great stuff on fri #sxsw: marissa mayer (google) tim o'reilly (tech books/conferences) &amp;amp; matt mullenweg (wordpress)
5     #sxsw is just starting #ctia is around the corner and #googleio is only a hop skip and a jump from there good time to be an #android
8     excited to meet the @samsungmobileus at #sxsw so i can show them my sprint galaxy s still running android 21   #fail
9     find &amp;amp; start impromptu parties at #sxsw with @hurricaneparty http://bitly/gvlrin (http://bitly/gvlrin) i can't wait til the andro
10    foursquare ups the game just in time for #sxsw http://jmp/grn7pk) (http://jmp/grn7pk)) - still prefer @gowalla by far best looking an
Name: text_processed, dtype: object
```

n [93]:
```
stop_words = stopwords.words('english')
stop_words.extend(['sxsw', 'sxswi', 'store', 'quot', 'mention', 'link', 'rt', 'amp', 'http', 'sxswrt', 'google', 'googles', 'app', 'apps',
```

n [94]:
```
# EDA word viz... for whole dataset.
# Import the wordcloud library
from wordcloud import WordCloud
# Join the different processed titles together.
long_string = ','.join(list(data4['text_processed'].values))
# Create a WordCloud object
wordcloud = WordCloud(background_color="white", stopwords=stop_words, max_words=200, height=400, width=600, contour_width=3, contour_color=
# Generate a word cloud
wordcloud.generate(long_string)
# Visualize the word cloud
wordcloud.to_image()
```

ut[94]:



n [58]:
```
# Prepare the data for LDA analysis
import gensim
from gensim.utils import simple_preprocess
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords

stop_words = stopwords.words('english')
stop_words.extend(['sxsw', 'sxswi', 'quot', 'mention', 'link', 'rt', 'amp', 'http', 'sxswrt', 'google', 'googles', 'app', 'apps', 'android'

def sent_to_words(sentences):
    for sentence in sentences:
        # deacc=True removes punctuations
        yield(gensim.utils.simple_preprocess(str(sentence), deacc=True))
def remove_stopwords(texts):
    return [[word for word in simple_preprocess(str(doc))
             if word not in stop_words] for doc in texts]
data = data4.text_processed.values.tolist()
data_words = list(sent_to_words(data))
# remove stop words
data_words = remove_stopwords(data_words)
print(data_words[:1][0][:30])
```

```
['sxtxstate', 'great', 'stuff', 'fri', 'marissa', 'mayer', 'tim', 'reilly', 'tech', 'books', 'conferences', 'matt', 'mullenweg', 'wordpress

[nltk_data] Downloading package stopwords to /Users/markp/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

n [59]:
```
# Create dictionary and corpus
import gensim.corpora as corpora
# Create Dictionary
id2word = corpora.Dictionary(data_words)
# Create Corpus
texts = data_words
# Term Document Frequency
corpus = [id2word.doc2bow(text) for text in texts]
# View
print(corpus[:1][0][:30])
```

```
[(0, 1), (1, 1), (2, 1), (3, 1), (4, 1), (5, 1), (6, 1), (7, 1), (8, 1), (9, 1), (10, 1), (11, 1), (12, 1), (13, 1)]
```

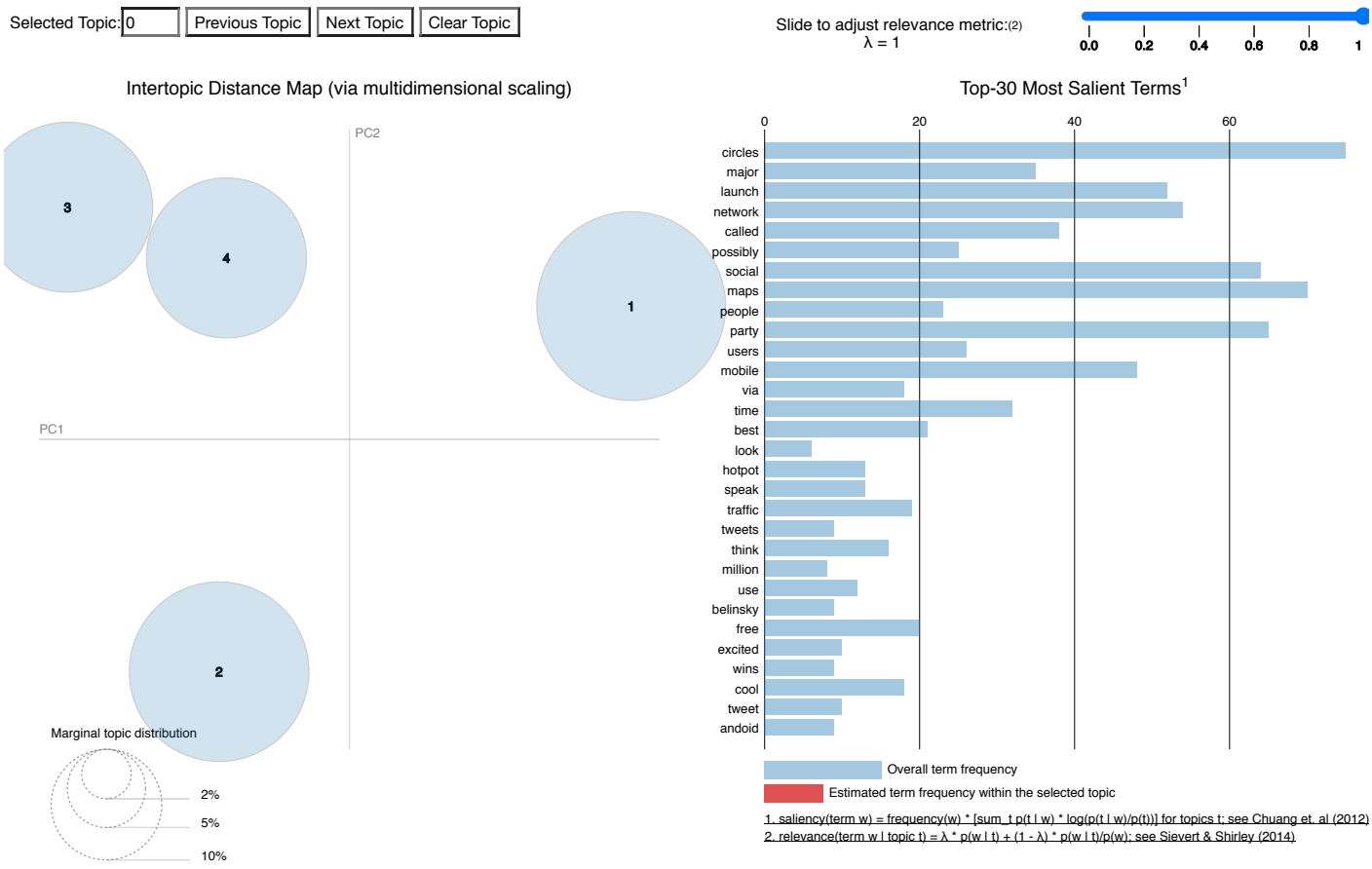
n [60]:
```
# Train the model
from pprint import pprint
# number of topics
num_topics = 4
# Build LDA model
lda_model = gensim.models.LdaMulticore(corpus=corpus,
                                       id2word=id2word,
                                       num_topics=num_topics)
# Print the Keyword in the 4 topics
pprint(lda_model.print_topics())
doc_lda = lda_model[corpus]
```

```
[(0,
  '0.032*"circles" + 0.025*"social" + 0.025*"network" + 0.024*"launch" + '
  '0.019*"called" + 0.019*"major" + 0.014*"possibly" + 0.009*"party" + '
  '0.008*"great" + 0.008*"via"'),
 (1,
  '0.010*"maps" + 0.009*"party" + 0.009*"mayer" + 0.009*"best" + '
  '0.008*"marissa" + 0.008*"mobile" + 0.007*"cool" + 0.006*"get" + '
  '0.006*"thanks" + 0.006*"great"'),
 (2,
  '0.010*"people" + 0.010*"marissa" + 0.010*"mobile" + 0.010*"mayer" + '
  '0.010*"maps" + 0.008*"around" + 0.008*"good" + 0.008*"traffic" + '
  '0.007*"think" + 0.006*"route"'),
 (3,
  '0.023*"maps" + 0.021*"party" + 0.013*"mobile" + 0.011*"time" + '
  '0.009*"users" + 0.008*"social" + 0.007*"circles" + 0.007*"great" + '
  '0.007*"check" + 0.006*"search"')]
```

n [61]:
```
# Visualize the resulting clusters
pyLDAvis.enable_notebook()

vis = pyLDAvis.gensim.prepare(lda_model, corpus, id2word)
vis
```

```
/Users/markp/opt/anaconda3/envs/learn-env/lib/python3.6/site-packages/pyLDAvis/_prepare.py:257: FutureWarning: Sorting because non-concaten
igned. A future version
of pandas will change to not sort by default.

To accept the future behavior, pass 'sort=False'.

To retain the current behavior and silence the warning, pass 'sort=True'.

  return pd.concat([default_term_info] + list(topic_dfs))
```

ut[61]:



4. Try seperate LDA for positive tweets (Apple)

This is data3 and has 1945 tweets.

n [62]:
```
data3.shape
```

ut[62]:
```
(1945, 4)
```

n [63]:
```
# TEXT Pre-processing
# Load the regular expression library
import re
# Remove punctuation
data3['text_processed'] = \
data3['text'].map(lambda x: re.sub('[,\.!?]', '', x))
# Convert the titles to lowercase
data3['text_processed'] = \
data3['text_processed'].map(lambda x: x.lower())
# Print out the first rows of papers
data3['text_processed'].head()
```

```
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<input>:5: DeprecationWarning: invalid escape sequence \.
<ipython-input-63-6bd0b8104945>:5: DeprecationWarning: invalid escape sequence \.
  data3['text_processed'] = data3['text'].map(lambda x: re.sub('[,\.!?]', '', x))
/Users/markp/opt/anaconda3/envs/learn-env/lib/python3.6/site-packages/ipykernel_launcher.py:5: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (
a.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  """
/Users/markp/opt/anaconda3/envs/learn-env/lib/python3.6/site-packages/ipykernel_launcher.py:7: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (
a.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  import sys
```

ut[63]:
```
1     @jessedee know about @fludapp  awesome ipad/iphone app that you'll likely appreciate for its design also they're giving free ts at #s
2     @swonderlin can not wait for #ipad 2 also they should sale them down at #sxsw
6     beautifully smart and simple idea rt @madebymany @thenextweb wrote about our #hollergram ipad app for #sxsw http://bitly/ieavob (http
7     counting down the days to #sxsw plus strong canadian dollar means stock up on apple gear
12    great #sxsw ipad app from @madebymany: http://tinyurlcom/4nqv92l (http://tinyurlcom/4nqv92l)
Name: text_processed, dtype: object
```

n [90]:
```
les', 'app', 'apps', 'android', 'austin', 'quotgoogle', 'new', 'today', 'one', 'apple', 'ipad', 'ipads', 'iphone', 'ipad2', 'apples', 'quot
```

n [92]:
```
# EDA word viz... for whole dataset.
# Import the wordcloud library
from wordcloud import WordCloud
# Join the different processed titles together.
long_string = ','.join(list(data3['text_processed'].values))
# Create a WordCloud object
wordcloud = WordCloud(background_color="white", stopwords=stop_words, max_words=200, height=400, width=600, contour_width=3, contour_color=
# Generate a word cloud
wordcloud.generate(long_string)
# Visualize the word cloud
wordcloud.to_image()
```

ut[92]:

n [69]:
```
# Prepare the data for LDA analysis
import gensim
from gensim.utils import simple_preprocess
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords

stop_words = stopwords.words('english')
stop_words.extend(['sxsw', 'sxswi', 'store', 'quot', 'mention', 'link', 'rt', 'amp', 'http', 'sxswrt', 'google', 'googles', 'app', 'apps',

def sent_to_words(sentences):
    for sentence in sentences:
        # deacc=True removes punctuations
        yield(gensim.utils.simple_preprocess(str(sentence), deacc=True))
def remove_stopwords(texts):
    return [[word for word in simple_preprocess(str(doc))
             if word not in stop_words] for doc in texts]
data = data3.text_processed.values.tolist()
data_words = list(sent_to_words(data))
# remove stop words
data_words = remove_stopwords(data_words)
print(data_words[:1][0][:30])
```

```
[nltk_data] Downloading package stopwords to /Users/markp/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!

['jessedee', 'know', 'fludapp', 'awesome', 'likely', 'appreciate', 'design', 'also', 'giving', 'free', 'ts']
```

n [70]:
```
# Create dictionary and corpus
import gensim.corpora as corpora
# Create Dictionary
id2word = corpora.Dictionary(data_words)
# Create Corpus
texts = data_words
# Term Document Frequency
corpus = [id2word.doc2bow(text) for text in texts]
# View
print(corpus[:1][0][:30])
```

```
[(0, 1), (1, 1), (2, 1), (3, 1), (4, 1), (5, 1), (6, 1), (7, 1), (8, 1), (9, 1), (10, 1)]
```

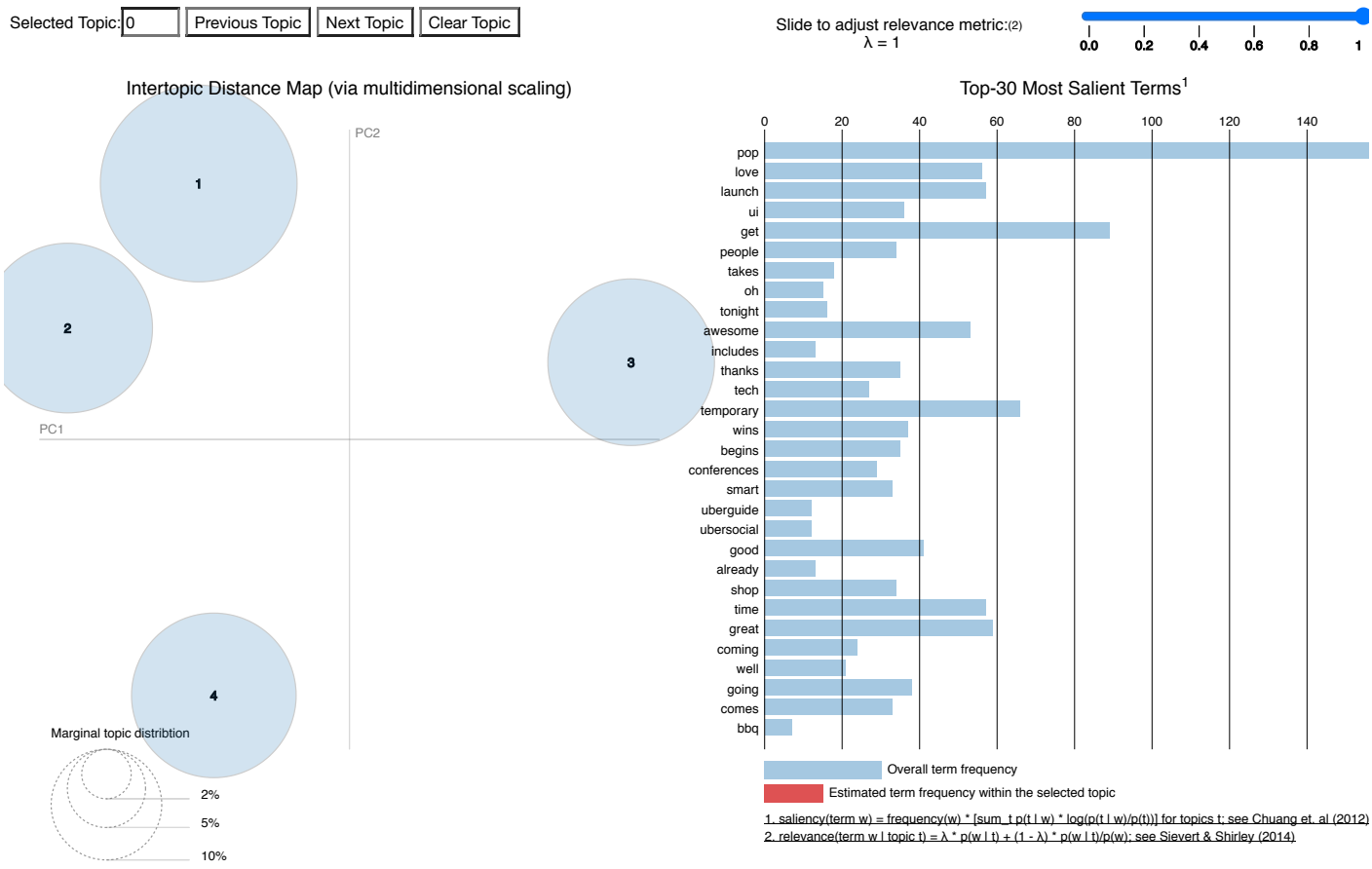
n [71]:
```
# Train the model
from pprint import pprint
# number of topics
num_topics = 4
# Build LDA model
lda_model = gensim.models.LdaMulticore(corpus=corpus,
                                       id2word=id2word,
                                       num_topics=num_topics)
# Print the Keyword in the 4 topics
pprint(lda_model.print_topics())
doc_lda = lda_model[corpus]
```

```
[(0,
  '0.011*"pop" + 0.009*"launch" + 0.007*"awesome" + 0.007*"line" + 0.007*"get" '
  '+ 0.006*"free" + 0.006*"people" + 0.006*"via" + 0.006*"cool" + '
  '0.005*"opening"'),
 (1,
  '0.011*"love" + 0.011*"get" + 0.008*"cool" + 0.007*"like" + 0.006*"go" + '
  '0.006*"good" + 0.006*"free" + 0.006*"popup" + 0.005*"ever" + '
  '0.005*"launch"'),
 (2,
  '0.009*"pop" + 0.008*"via" + 0.008*"temporary" + 0.008*"downtown" + '
  '0.007*"ui" + 0.007*"open" + 0.007*"get" + 0.007*"even" + 0.007*"line" + '
  '0.006*"opening"'),
 (3,
  '0.025*"pop" + 0.008*"great" + 0.008*"line" + 0.008*"time" + 0.006*"video" + '
  '0.006*"day" + 0.006*"going" + 0.005*"smart" + 0.005*"awesome" + '
  '0.005*"shop"')]
```

n [72]:
```
# Visualize the resulting clusters
pyLDAvis.enable_notebook()

vis = pyLDAvis.gensim.prepare(lda_model, corpus, id2word)
vis
```

```
/Users/markp/opt/anaconda3/envs/learn-env/lib/python3.6/site-packages/pyLDAvis/_prepare.py:257: FutureWarning: Sorting because non-concaten
igned. A future version
of pandas will change to not sort by default.

To accept the future behavior, pass 'sort=False'.

To retain the current behavior and silence the warning, pass 'sort=True'.

  return pd.concat([default_term_info] + list(topic_dfs))
```

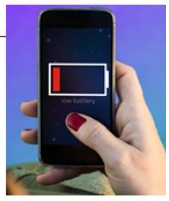
ut[72]:



Observations on LDA process

So we can see that the LDA helps us in triangulating some of the key issues from the negative tweets. Combined with the word clouds (word c
the model's feature importance, and from reading a sample of the tweets, we can make some recommendations to Apple and Google. See the summ
recommendations below (taken from the presentation).

Recommendations



GOOGLE

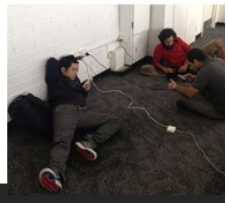
- Google Circles: too much competition
- Google Maps: fix navigation errors quicker
- Google Search: at risk from Microsoft Bing
- Company: "focus on things that matter"

APPLE

- iPhone: battery lacks longevity
- News: needs improvement and innovation
- iPad2 Launch: long lines and poor CS
- Conference: more phone charging stations

Process Improvements

- **Dashboard Approach:** model tuned for speed and frequent reporting
- **Tweets:** may not be the best source of actionable feedback (part of a mix)



In [ ]:

APPENDIX

In [ ]:
```
# LDA from Yish study group


import numpy as np
import warnings
warnings.filterwarnings('ignore')

%matplotlib inline
import matplotlib.pyplot as plt
plt.style.use('ggplot')

from sklearn.decomposition import PCA

import gensim.downloader as api
from gensim.test.utils import datapath
from gensim.models import KeyedVectors
```

In [ ]:
```
word_vectors = api.load("glove-wiki-gigaword-100")
```

In [ ]:
```
word_vectors.most_similar('coffee')
```

In [ ]:

In [ ]:
```
# TOPIC MODELING

import gensim

from nltk.corpus import stopwords
import gensim.corpora as corpora

import pyLDAvis.gensim
pyLDAvis.enable_notebook()
```

In [ ]:
```
def process_words(texts, stop_words=stopwords.words("english"), allowed_postags=['NOUN', 'ADJ', 'VERB', 'ADV']):

    texts = [[word for word in doc.split() if word not in stop_words] for doc in texts]
    texts_out = []
    nlp = spacy.load('en_core_web_sm', disable=['parser', 'ner'])
    for sent in texts:
        doc = nlp(" ".join(sent))
        texts_out.append([token.lemma_ for token in doc if token.pos_ in allowed_postags])
    # remove stopwords once more after lemmatization
    texts_out = [[word for word in doc if word not in stop_words] for doc in texts_out]
    return texts_out

data_ready = process_words(raw.body)
```

In [ ]:
```
# Create Dictionary
id2word = corpora.Dictionary(data_ready)

# Create Corpus: Term Document Frequency
corpus = [id2word.doc2bow(text) for text in data_ready]

# Build LDA model
lda_model = gensim.models.ldamodel.LdaModel(corpus=corpus,
                                            id2word=id2word,
                                            num_topics=4,
                                            random_state=100,
                                            update_every=1,
                                            chunksize=10,
                                            passes=10,
                                            alpha='symmetric',
                                            iterations=100,
                                            per_word_topics=True)
```

In [ ]:
```
print(lda_model.print_topics())
```

In [ ]:
```
vis = pyLDAvis.gensim.prepare(lda_model, corpus, dictionary=lda_model.id2word)
vis
```

In [ ]:
```
# Analyzing the LDA results - this is from the article used above, but I'm unsure of the saving step and path designation.
import pyLDAvis.gensim
import pickle
import pyLDAvis

# Visualize the topics
pyLDAvis.enable_notebook()
LDAvis_data_filepath = os.path.join('./results/ldavis_prepared_'+str(num_topics))
# # this is a bit time consuming - make the if statement True
# # if you want to execute visualization prep yourself
if 1 == 1:
    LDAvis_prepared = pyLDAvis.gensim.prepare(lda_model, corpus, id2word)
    with open(LDAvis_data_filepath, 'wb') as f:
        pickle.dump(LDAvis_prepared, f)
# load the pre-prepared pyLDAvis data from disk
with open(LDAvis_data_filepath, 'rb') as f:
    LDAvis_prepared = pickle.load(f)
pyLDAvis.save_html(LDAvis_prepared, './results/ldavis_prepared_'+ str(num_topics) +'.html')
LDAvis_prepared
```

In [ ]: