Reactions per turn specific to politics of app user

Continuing to other task than predicting total reactions in general.

In [1]:
```
import pandas as pd
import reactions
import nltk
import random
import matplotlib.pyplot as plt
from pandas.tools.plotting import scatter_matrix
```

In [2]:
```
%time r = reactions.link_reactions_to_transcript('data/reactions_oct3_4project.csv','corpora/oct3_coded_transcript_sync.csv')
```

```
CPU times: user 8.28 s, sys: 0.51 s, total: 8.79 s
Wall time: 8.80 s
```

In [3]:
```
r2 = r.copy()
del r2["Sync'd start"]
del r2["Sync'd end"]
del r2["Time"]
del r2["Speaker"]
r2.head(2)
```

Out[3]:

	Frame	QuestionTopic	Reaction_what	Reaction_who	Tone	Topic	Transcript	UserID	start	stateme
0	9	99	Agree	Moderator	0	9	Good evening from the Magness Arena at the Uni...	ag1zfnJlYWN0bGFicy00ciwLEgRVc2VyIiJhX2YzNTQxZW...	01:02:01	0
56861	9	99	Disagree	Moderator	0	9	Good evening from the Magness Arena at the Uni...	ag1zfnJlYWN0bGFicy01ciwLEgRVc2VyIiJhX2U3YmFkZT...	01:02:01.401000	0

Political questionnaire data

In [4]:
```
%time p = reactions.split_reactions_file('data/reactions_oct3_4project.csv')['quest_political']
```

```
CPU times: user 4.46 s, sys: 0.32 s, total: 4.79 s
Wall time: 4.79 s
```

In [5]:
```
p2 = p[['UserID','party_1','political_views_2','candidate_choice_3','confidence_in_choice_4','likely_to_vote_5','candidate_preferred_29']]
p2.head(2)
```

Out[5]:

	UserID	party_1	political_views_2	candidate_choice_3	confidence_in_choice_4	likely_to_vote_5	candidate_pref
0	ag1zfnJlYWN0bGFicy00ciwLEgRVc2VyIiJhX2E0Mjc1MD...	closest to republican party	73	romney	100	100	NaN
62	ag1zfnJlYWN0bGFicy00ciwLEgRVc2VyIiJhX2E0Mzk5OD...	closest to democratic party	20	obama	100	100	NaN

In [6]:
```
p2
```

Out[6]:
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3767 entries, 0 to 193268
Data columns:
UserID                  3767  non-null values
party_1                 3733  non-null values
political_views_2       3733  non-null values
candidate_choice_3      3733  non-null values
confidence_in_choice_4  3733  non-null values
likely_to_vote_5        3733  non-null values
candidate_preferred_29  2118  non-null values
dtypes: float64(4), object(3)
```

There are ~30 users for whom we don't have political preference info, and the the candidate_preferred_29 col was often left blank.

Simplify party membership into R/D/oth

Let's group the users into D/R/other.

In [7]:
```
p2.groupby('party_1').agg('count').UserID
```

Out[7]:
```
party_1
closest to democratic party    1267
closest to republican party     479
independent                     598
lean democrat                   781
lean republican                 527
no answer                        81
Name: UserID
```

In [8]:
```
p2['party'] = p2.party_1.apply(lambda a: {'closest to democratic party':'democrat',
                                          'lean democrat':'democrat',
                                          'lean republican':'republican',
                                          'closest to republican party':'republican'}.get(a,'other'))
p2.groupby('party').agg('count').UserID
```

Out[8]:
```
party
democrat      2048
other          713
republican    1006
Name: UserID
```

Merge political questionnaire with reactions

In [9]:
```
%time r3 = r2.merge(p2[['UserID','party']])
print 'pre-merge:',len(r2),'post-merge:',len(r3)
r3.head(2)
```

```
CPU times: user 0.51 s, sys: 0.04 s, total: 0.55 s
Wall time: 0.55 s
pre-merge: 189015 post-merge: 189015
```

Out[9]:

	Frame	QuestionTopic	Reaction_what	Reaction_who	Tone	Topic	Transcript	UserID	start	statement	t
0	9	99	Agree	Moderator	0	9	Good evening from the Magness Arena at the Uni...	ag1zfnJlYWN0bGFicy00ciwLEgRVc2VyIiJhX2YzNTQxZW...	01:02:01	0	1
1	3	5	Dodge	Obama	1	5	Over the last 30 months, we've seen 5 million ...	ag1zfnJlYWN0bGFicy00ciwLEgRVc2VyIiJhX2YzNTQxZW...	01:05:34.890000	30	2

Group by turn

Statements

In [11]:
```
st = r3.groupby(['statement']).first()[['Speaker_name','Transcript','turn']]
st.head(2)
```

Out[11]:

	Speaker_name	Transcript	turn
statement			
0	Moderator	Good evening from the Magness Arena at the Uni...	1
1	Moderator	I'm Jim Lehrer of the PBS NewsHour,	1

Turns

In [12]:
```
t = pd.DataFrame({'speaker':st.groupby('turn').first().Speaker_name,
                  'reactions':r3.groupby('turn').count().Speaker_name,
                  'statements':st.groupby('turn').count().turn,
                  'text':st.groupby('turn').apply(lambda x: ''.join(x.Transcript)),
                  'agree':r3[r3.Reaction_what=='Agree'].groupby('turn').count().turn,
                  'agree_dem':r3[(r3.party=='democrat') & (r3.Reaction_what=='Agree')].groupby('turn').count().turn,
                  'agree_rep':r3[(r3.party=='republican') & (r3.Reaction_what=='Agree')].groupby('turn').count().turn,
                  'disagree':r3[r3.Reaction_what=='Disagree'].groupby('turn').count().turn,
                  'disagree_dem':r3[(r3.party=='democrat') & (r3.Reaction_what=='Disagree')].groupby('turn').count().turn,
                  'disagree_rep':r3[(r3.party=='republican') & (r3.Reaction_what=='Disagree')].groupby('turn').count().turn,
                  })
t['words'] = t.text.apply(lambda txt: [t.lower() for t in nltk.tokenize.word_tokenize(txt) if t.isalpha()])
t['word_count'] = t.words.apply(lambda words: len(words))
t['r_per_st'] = 1.0 * t.reactions / t.statements
t['r_per_w'] = 1.0 * t.reactions / t.word_count
t['a_to_d_dems'] = t.agree_dem / t.disagree_dem
```

```
t['a_to_d_reps'] = t.agree_rep / t.disagree_rep
ranked_unigrams = nltk.FreqDist([w for word_list in t.words for w in word_list]).keys()
MAX_FEATURES = 900
t['unigrams'] = t.words.apply(lambda words: {w:True for w in words if w in ranked_unigrams[:MAX_FEATURES]})
t['unigram_count'] = t.unigrams.apply(lambda unigrams: len(unigrams))
```

In [13]:
```
t.head(2)
```

Out[13]:

	agree	agree_dem	agree_rep	disagree	disagree_dem	disagree_rep	reactions	speaker	statements	text	words	word_count	r_per_st	r_per_w	a_to
turn															
1	488	313	94	161	54	44	812	Moderator	20	Good evening from the Magness Arena at the Uni...	[good, evening, from, the, magness, arena, at,...	257	40.6	3.159533	5.79
2	2678	1958	273	460	85	299	4213	Obama	22	Well, thank you very much, Jim, for this oppor...	[well, thank, you, very, much, jim, for, this,...	278	191.5	15.154676	23.0

Filter

For now, we get rid of the really short turns, which would seem to likely have noise from adjacent turns and the small numbers of words make the math more sketchy.

In [14]:
```
MIN_WORDS = 30
t2 = t[t.word_count >= MIN_WORDS]
print len(t),'->',len(t2)
```

```
190 -> 71
```

What to predict?

Agree - to - disagree ratio

At least one person of each party agrees and disagrees for each turn.

In [15]:
```
t.describe()
```

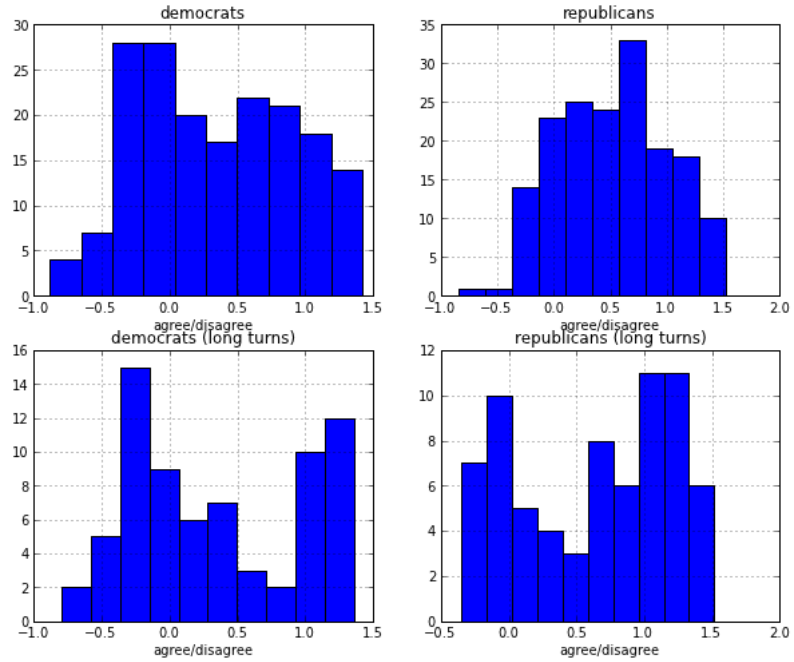
Out[15]:

	agree	agree_dem	agree_rep	disagree	disagree_dem	disagree_rep	reactions	statements	word_count	r_per_st	r_per_w	a_to_d_dems
count	187.000000	185.000000	184.000000	188.000000	180.000000	170.000000	190.000000	190.000000	190.000000	190.000000	190.000000	179.000000
mean	577.540107	322.135135	162.570652	226.239362	150.044444	55.376471	994.815789	6.168421	77.452632	111.787644	inf	4.667936
std	983.398743	658.687742	305.054748	392.908027	309.975595	114.853287	1622.915791	8.907107	121.482315	81.179389	NaN	5.658037
min	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000	1.000000	0.166667	0.130435
25%	29.000000	16.000000	9.000000	12.000000	8.000000	3.000000	53.500000	1.000000	5.000000	43.125000	9.045092	0.750000
50%	83.000000	42.000000	21.500000	38.500000	24.500000	9.500000	141.500000	2.000000	14.000000	92.000000	12.493183	2.000000
75%	564.500000	247.000000	177.250000	284.250000	124.000000	33.750000	1075.500000	6.750000	84.000000	167.028846	15.500000	6.375000
max	4953.000000	3588.000000	1928.000000	2572.000000	2125.000000	580.000000	7777.000000	54.000000	497.000000	393.500000	inf	26.333333

Most of the time more people seem to be agreeing than disagreeing. Republicans especially so..

In [84]:
```
figsize(10,8)

subplot(221)
log10(t.a_to_d_dems).hist()
xlabel('agree/disagree')
title('democrats')

subplot(222)
log10(t.a_to_d_reps).hist()
xlabel('agree/disagree')
title('republicans')

subplot(223)
log10(t2.a_to_d_dems).hist()
xlabel('agree/disagree')
title('democrats (long turns)')

subplot(224)
log10(t2.a_to_d_reps).hist()
```

```
xlabel('agree/disagree')
title('republicans (long turns)')

show()
```



So it seems that taking out reactions from short turns reveals more polarization. Perhaps this is because on short turns there is just more noise from adjacent turns? Or that people become more and more energized in their responses as the speakers continue to talk?

In [94]:
```
PERC = .03
print '{:_^80}'.format('dems agree'.upper())
for v in t2[t2.a_to_d_dems > t2.a_to_d_dems.quantile(1-PERC)].text.values: print v+'\n'
print '{:_^80}'.format('dems disagree'.upper())
for v in t2[t2.a_to_d_dems < t2.a_to_d_dems.quantile(PERC)].text.values: print v+'\n'
```

```
_____________________________________DEMS AGREE_____________________________________
Well, thank you very much, Jim, for this opportunity.I want to thank Governor Romney and the University of Denver for your hospitality.There are a lot of points that I want to make tonightbut the most important one is that 20 years ago I became the luckiest man on earth because Michelle Obama agreed to marry me.And so I just want to wish, Sweetie, you happy anniversary and let you know that a year from now, we will not be celebrating it in front of 40 million people.You know, four years ago we went through the worst financial crisis since the Great Depression.Millions of jobs were lost.The auto industry was on the brink of collapse.The financial system had frozen up.And because of the resilience and the determination of the American people, we've begun to fight our way back.Over the last 30 months, we've seen 5 million jobs in the private sector created.The auto industry has come roaring backand housing has begun to rise.But we all know that we've still got a lot of work to do.And so the question here tonight is not where we've been but where we're going.Governor Romney has a perspective that says if we cut taxes, skewed towards the wealthy, and roll back regulations that we'll be better off.I've got a different view.I think we've got to invest in education and training.I think it's important for us to develop new sources of energy here in Americathat we change our tax code to make sure that we're helping small businesses and companies that are investing here in the United States,that we take some of the money that we're saving as we wind down two wars to rebuild Americaand that we reduce our deficit in a balanced way that allows us to make these critical investments.

Well, for 18 months he's been running on this tax plan.And now, five weeks before the election, he's saying that his big, bold idea is
never mind.And the fact is that if you are lowering the rates the way you describe, Governor, then it is not possible to come up with enough deductions and loopholes that only affect high-income individuals to avoid either raising the deficit or burdening the middle class.It's -- it's math.It's arithmetic.Now, Governor Romney and I do share a deep interest in encouraging small-business growth.So at the same time that my tax plan has already lowered taxes for 98 percent of families, I also lowered taxes for small businesses 18 times.And what I want to do is continue the tax rates -- the tax cuts that we put into place for small businesses and families.But I have said that for incomes over $250,000 a year that we should go back to the rates that we had when Bill Clinton was president, when we created 23 million new jobs, went from deficit to surplus and created a whole lot of millionaires to boot.And the reason this is important is because by doing that, we can not only reduce the deficit, we can not only encourage job growth through small businessesbut we're also able to make the investments that are necessary in education or in energy.And we do have a difference, though, when it comes to definitions of small business.Now, under -- under my plan, 97 percent of small businesses would not see their income taxes go upGovernor Romney says, well, those top 3 percent, they're the job creators.They'd be burdened.But under Governor Romney's definition, there are a whole bunch of millionaires and billionaires who are small businesses.Donald Trump is a small business.And I know Donald Trump doesn't like to think of himself as small anythingbut that's how you define small businesses if you're getting business income.And that kind of approach, I believe, will not grow our economy because the only way to pay for it without either burdening the middle class or blowing up our deficit is to make drastic cuts in things like education, making sure that we are continuing to invest in basic science and research, all the things that are helping America grow.And I think that would be a mistake.

You've got to have --If we're serious, we've got to take a balanced, responsible approach.And by the way, this is not just when it comes to individual taxes.Let's talk about corporate taxes.Now, I've identified areas where we can, right away, make a change that I believe would actually help the economy.The -- the oil industry gets $4 billion a year in corporate welfare.Basically, they get deductions that those small businesses that Governor Romney refers to, they don't get.Now, does anybody think that ExxonMobil needs some extra money when they're making money every time you go to the pump?Why wouldn't we want to eliminate that?Why wouldn't we eliminate tax breaks for corporate jets? My attitude is if you got a corporate jet, you can probably afford to pay full freight, not get a special break for it.When it comes to corporate taxes, Governor Romney has said he wants to, in a revenue-neutral way, close loopholes, deductions -- he hasn't identified which ones they are -- but thereby bring down the corporate rate.Well, I want to do the same thing, but I've actually identified how we can do that.And part of the way to do it is to not give tax breaks to companies that are shipping jobs overseas.Right now you can actually take a deduction for moving a plant overseas.I think most Americans would say that doesn't make sense.And all that raises revenue.And so if we take a balanced approach, what that then allows us to do is also to help young people, the way we already have during my administration, make sure that they can afford to go to college.It means that the teacher that I met in Las Vegas, wonderful young lady, who describes to me -- she's got 42 kids in her class.The first two weeks, she's got them -- some of them sitting on the floor until finally they get reassigned.They're using textbooks that are 10 years old.That is not a recipe for growth: that's not how America was built.And so budgets
```

reflect choices.Ultimately we're going to have to make some decisions.And if we're asking for no revenue, then that means that we've got to get rid of a whole bunch of stuffand the magnitude of the tax cuts that you're talking about, Governor, would end up resulting in severe hardship for people but more importantly, would not help us grow.As I indicated before, when you talk about shifting Medicaid to states, we're talking about potentially a -- a 30 -- a 30 percent cut in Medicaid over time.Now, you know, that may not seem like a big deal when it just is -- you know, numbers on a sheet of paperbut if we're talking about a family who's got an autistic kid and is depending on that Medicaid, that's a big problem.And governors are creative.There's no doubt about that.But they're not creatie enough to make up for the 30 percent revenue on something like Medicaid.What ends up happening is some people end up not getting help.

_____DEMS DISAGREE_____
Let me -- let me repeat -- let me repeat what I said -- (inaudible).I'm not in favor of a $5 trillion tax cut.That's not my plan.My plan is not to put in place any tax cut that will add to the deficit.That's point one.So you may keep referring to it as a $5 trillion tax cut, but that's not my plan.

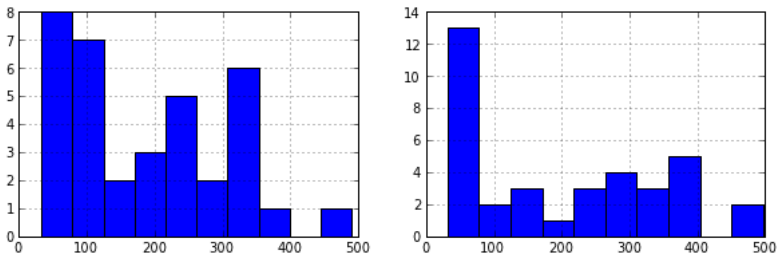
I sure do.Well, in part, it comes, again, from my experience.I was in New Hampshire.A woman came to me, and she said, look, I can't afford insurance for myself or my son.I met a couple in Appleton, Wisconsin, and they said, we're thinking of dropping our insurance; we can't afford it.And the number of small businesses I've gone to that are saying they're dropping insurance because they can't afford it -- the cost of health care is just prohibitive.And -- and we've got to deal with cost.And unfortunately, when -- when you look at ObamacareThe Congressional Budget Office has said it will cost $2,500 a year more than traditional insurance.So it's adding to cost.And as a matter of fact, when the president ran for office, he said that by this year he would have brought down the cost of insurance for each family by $2,500 a family.Instead, it's gone up by that amount.So it's expensive.Expensive things hurt families.So that's one reason I don't want it.Second reason, it cuts $716 billion from Medicare to pay for it.I want to put that money back in Medicare for our seniors.Number three, it puts in place an unelected board that's going to tell people, ultimately, what kind of treatments they can have.I don't like that idea.Fourth, there was a survey done of small businesses across the country.It said, what's been the effect of Obamacare on your hiring plans?And three-quarters of them said, it makes us less likely to hire people.I just don't know how the president could have come into office, facing 23 million people out of work, rising unemployment, an economic crisis at the -- at the kitchen table and spent his energy and passion for two years fighting for Obamacare instead of fighting for jobs for the American people.It has killed jobs and the best course for health care is to do what we did in my state, craft a plan at the state level that fits the needs of the state.And then let's focus on getting the costs down for people rather than raising it with the $2,500 additional premium.

If I'm elected, we won't have Obamacare. We'll put in place the kind of principles that I put in place in my own state and allow each state to craft their own programs to get people insured. And we'll focus on getting the cost of health care down.If the president were to be re-elected, you're going to see a $716 billion cut to Medicare. You'll have 4 million people who will lose Medicare advantage. You'll have hospitals and providers that'll no longer accept Medicare patients.I'll restore that $716 billion to Medicare.And finally, military. If the president's re-elected, you'll see dramatic cuts to our military. The secretary of defense has said these would be even devastating. I will not cut our commitment to our military. I will keep America strong and get America's middle class working again.Thank you, Jim.

How large are the turns where dems either agree or disagree?

In [96]:
```
figsize(10,3)
subplot(121)
t2[t2.a_to_d_dems < t2.a_to_d_dems.quantile(.5)].word_count.hist()
subplot(122)
t2[t2.a_to_d_dems >= t2.a_to_d_dems.quantile(.5)].word_count.hist()
```

Out[96]:
```
<matplotlib.axes.AxesSubplot at 0x9a9a3b0>
```



The threshold for converting the ratio to a true/false label is ~1.7 dems agreeing to 1 dem disagreeing.

In [16]:
```
t2['label'] = t2.a_to_d_dems >= t2.a_to_d_dems.quantile(.5)
print t2.a_to_d_dems.quantile(.5)
t2.label.describe()
```

```
1.7027027027
```

Out[16]:
```
count           71
mean     0.5070423
std      0.5035088
min          False
25%              0
50%              1
75%              1
max           True
```

Train and test experiment on dems

In [17]:
```
ex = t2
```

In [18]:
```
train_rows = random.sample(ex.index, len(ex)*9/10)
trn = ex.ix[train_rows]
tst = ex.drop(train_rows)
print len(trn)
print len(tst)
```

```
63
8
```

In [19]:
```
%time cl = nltk.NaiveBayesClassifier.train(zip(trn.unigrams, trn.label))
```

```
CPU times: user 0.07 s, sys: 0.01 s, total: 0.09 s
Wall time: 0.08 s
```

In [20]:
```
nltk.classify.accuracy(cl, zip(tst.unigrams, tst.label))
```

Out[20]:
```
0.75
```

In [25]:
```
cl.show_most_informative_features(10)
```

```
Most Informative Features
                  romney = True             True : False  =     14.5 : 1.0
                 america = True            False : True   =      7.9 : 1.0
                     get = True            False : True   =      6.6 : 1.0
                  course = True            False : True   =      5.8 : 1.0
                   comes = True             True : False  =      5.5 : 1.0
                  reason = True             True : False  =      4.8 : 1.0
               companies = True             True : False  =      4.8 : 1.0
                 problem = True             True : False  =      4.8 : 1.0
              difference = True             True : False  =      4.8 : 1.0
                governor = True             True : False  =      4.8 : 1.0
```

Whoa! Dems really hate america.. haha

Train and test experiment on reps

The threshold we will use for republicans is higher than the threshold for democrats. This appears to be because more republicans were agreeing with what was said during the debate over all compared to democrats.

In [28]:
```
t2['label2'] = t2.a_to_d_reps >= t2.a_to_d_reps.quantile(.5)
print t2.a_to_d_reps.quantile(.5)
t2.label2.describe()
```

```
5.23076923077
```

Out[28]:
```
count           71
mean     0.5070423
std      0.5035088
min          False
25%              0
50%              1
75%              1
max           True
```

In [29]:
```
ex = t2
```

In [30]:
```
train_rows2 = random.sample(ex.index, len(ex)*9/10)
trn2 = ex.ix[train_rows2]
tst2 = ex.drop(train_rows2)
print len(trn2)
print len(tst2)
```

```
63
8
```

In [31]:
```
%time cl2 = nltk.NaiveBayesClassifier.train(zip(trn2.unigrams, trn2.label2))
```

```
CPU times: user 0.06 s, sys: 0.01 s, total: 0.08 s
Wall time: 0.07 s
```

In [32]:
```
nltk.classify.accuracy(cl2, zip(tst2.unigrams, tst2.label2))
```

Out[32]:
```
0.75
```

In [33]:
```
cl2.show_most_informative_features(10)
```

```
Most Informative Features
                  system = True            False : True   =      6.9 : 1.0
                approach = True            False : True   =      6.2 : 1.0
                  romney = True            False : True   =      6.1 : 1.0
                   comes = True            False : True   =      5.5 : 1.0
              difference = True            False : True   =      5.5 : 1.0
                  before = True            False : True   =      5.5 : 1.0
                governor = True            False : True   =      5.2 : 1.0
                    only = True            False : True   =      4.8 : 1.0
                     top = True            False : True   =      4.8 : 1.0
                 america = True             True : False  =      4.6 : 1.0
```

Really, these train/test sets are **so** small, that (1) we can't draw much information from them without cross validation and (2) we are very prone to overfitting.

Hyperparams grid search on dems

Let's see if we can tune the max features hyper parameter (how many of the most frequent unigrams to use as features).

In [108]:
```
gr = t2.copy()
```

In [109]:
```
len(ranked_unigrams)
```

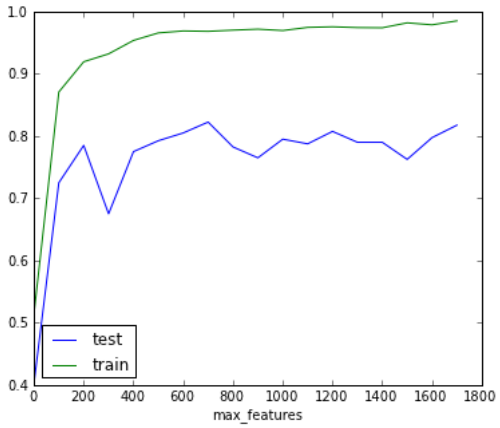
Out[109]: 1757

In [111]:
```
p = []
trn_means = []
tst_means = []
for max_feats in range(1,len(ranked_unigrams),100):
    gr['unigrams'] = gr.words.apply(lambda words: {w:True for w in words if w in ranked_unigrams[:max_feats]})
    trn_ac = []
    tst_ac = []
    print max_feats,
    for i in range(50):
        print i,
        train_rows = random.sample(gr.index, len(gr)*9/10)
        trn,tst = gr.ix[train_rows],gr.drop(train_rows)
        cl = nltk.NaiveBayesClassifier.train(zip(trn.unigrams, trn.label))
        trn_ac.append(nltk.classify.accuracy(cl, zip(trn.unigrams, trn.label)))
        tst_ac.append(nltk.classify.accuracy(cl, zip(tst.unigrams, tst.label)))
    p.append(max_feats)
    trn_means.append(mean(trn_ac))
    tst_means.append(mean(tst_ac))
    print ''
```

```
1 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48
49
101 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47
48 49
201 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47
48 49
301 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47
48 49
401 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47
48 49
501 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47
48 49
601 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47
48 49
701 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47
48 49
801 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47
48 49
901 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47
48 49
1001 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47
48 49
1101 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47
48 49
1201 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47
48 49
1301 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47
48 49
1401 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47
48 49
1501 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47
48 49
1601 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47
48 49
1701 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47
48 49
```

In [112]:
```
figsize(6,5)
results = pd.DataFrame({'max_features':p, 'train':trn_means, 'test':tst_means})
results.plot(x='max_features')
```

Out[112]: <matplotlib.axes.AxesSubplot at 0x9bc60b0>



In [113]:
```
results
```

Out[113]:

	max_features	test	train
0	1	0.4025	0.514286

1	101	0.7250	0.870794
2	201	0.7850	0.919365
3	301	0.6750	0.932063
4	401	0.7750	0.953651
5	501	0.7925	0.965714
6	601	0.8050	0.968889
7	701	0.8225	0.968254
8	801	0.7825	0.970159
9	901	0.7650	0.971746
10	1001	0.7950	0.969524
11	1101	0.7875	0.974603
12	1201	0.8075	0.975556
13	1301	0.7900	0.974286
14	1401	0.7900	0.973968
15	1501	0.7625	0.981905
16	1601	0.7975	0.978730
17	1701	0.8175	0.985079

It looks like going past ~700 unigram features is not helpful, and by then we are overfitting on train.

In [ ]: