Looking at reactions per-turn

We are analyzing the reactions linked to the debate transcript corpus on a turn-by-turn basis.

We have the over-all goal of predicting the amount of reaction to what a speaker is saying during his turn based on what words he uses.

In [141]:

```
import pandas as pd
import reactions
import nltk
import random
import matplotlib.pyplot as plt
from pandas.tools.plotting import scatter_matrix
```

Load the table.

In [2]:

```
%time t = reactions.link_reactions_to_transcript('data/reactions_oct3_4project.csv','corpora/oct3_coded_transcript_sync.csv')
t
```

```
CPU times: user 7.73 s, sys: 0.50 s, total: 8.23 s
Wall time: 8.30 s
```

Out[2]:

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 189015 entries, 0 to 191634
Data columns:
Frame            189015  non-null values
QuestionTopic    189015  non-null values
Reaction_what    189015  non-null values
Reaction_who     189015  non-null values
Speaker          189015  non-null values
Sync'd end       189015  non-null values
Sync'd start     189015  non-null values
Time             189015  non-null values
Tone             189015  non-null values
Topic            189015  non-null values
Transcript       189015  non-null values
UserID           189015  non-null values
start            189015  non-null values
statement        189015  non-null values
turn             189015  non-null values
dtypes: float64(6), int64(1), object(8)
```

In [3]:

```
t[['turn','Speaker','Transcript','start','Reaction_what','Reaction_who']].head(2)
```

Out[3]:

	turn	Speaker	Transcript	start	Reaction_what	Reaction_who
0	1	0	Good evening from the Magness Arena at the Uni...	01:02:01	Agree	Moderator
56861	1	0	Good evening from the Magness Arena at the Uni...	01:02:01.401000	Disagree	Moderator

In [4]:

```
t[['turn','Speaker','Transcript','start','Reaction_what','Reaction_who']].tail(2)
```

Out[4]:

	turn	Speaker	Transcript	start	Reaction_what	Reaction_who
68397	190	0	Thank you, and good night.	02:32:59.726000	Disagree	Romney
191634	190	0	Thank you, and good night.	02:32:59.840000	Agree	Romney

In [4]:

```
print t[:1]
```

```
   Frame  QuestionTopic Reaction_what Reaction_who  Speaker Sync'd end Sync'd start  \
0      9             99         Agree    Moderator        0    1:02:06      1:02:01   

                        Time  Tone  Topic  \
0  2012-10-04 01:02:00.967000     0      9   

                                          Transcript  \
0  Good evening from the Magness Arena at the Uni...   

                                              UserID     start  statement  turn  
0  ag1zfnJlYWN0bGFicy00ciwLEgRVc2VyIiJhX2YzNTQxZW...  01:02:01          0     1  
```

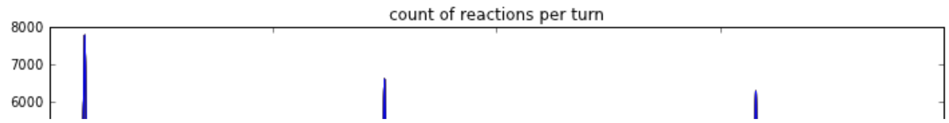
Number of reactions for each turn

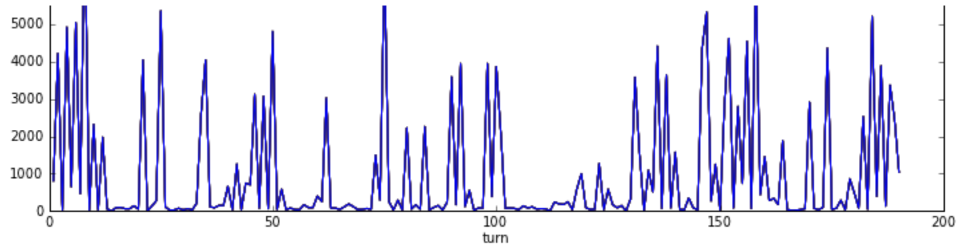
In [5]:

```
t.groupby('turn').count().plot(legend=False, figsize=(12, 4), title='count of reactions per turn')
```

Out[5]:

```
<matplotlib.axes.AxesSubplot at 0xa3aa690>
```



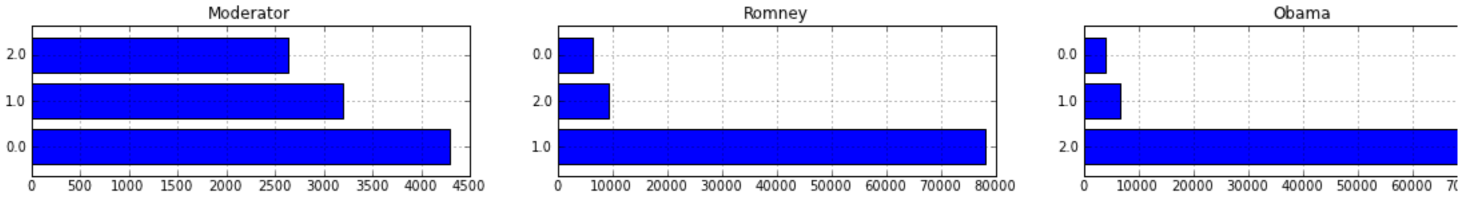


Looking at all reactions for each speaker

How does Speaker map to Reaction_who?

0 = Moderator **1** = Romney **2** = Obama

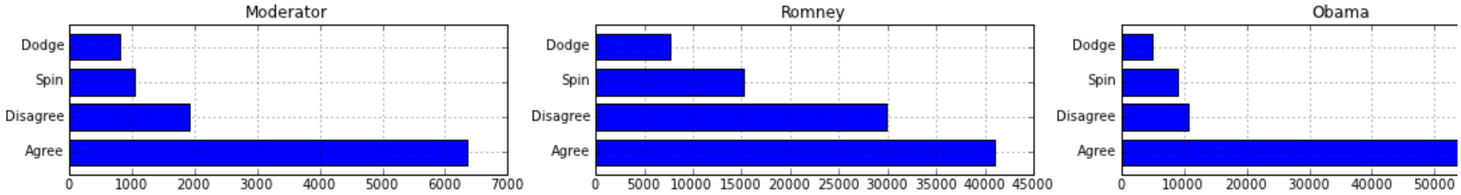
In [154]:
```
for i,s in enumerate(['Moderator','Romney','Obama']):
    plt.subplot(131+i)
    t[t.Reaction_who == s].Speaker.value_counts().plot(title=s, kind='barh')
plt.figsize(15,2)
plt.show()
```



It is interesting that a lot of the time people are reacting to people who are not speaking. Why is this? This is especially true when the moderator is speaking, but perhaps that is not unexpected.

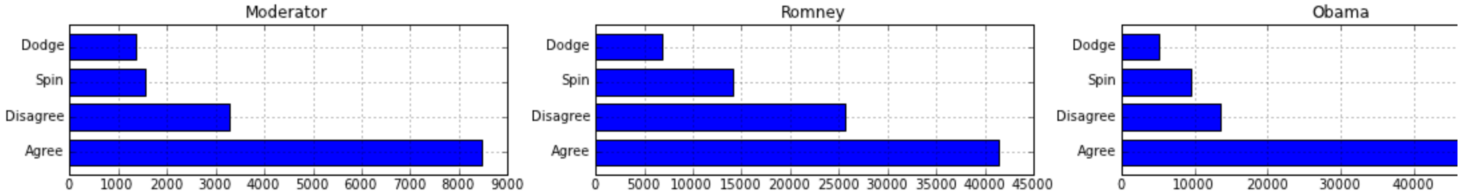
Now let's look at the the reaction data alone to see how people feel about each candidate.

In [153]:
```
for i,s in enumerate(['Moderator','Romney','Obama']):
    plt.subplot(131+i)
    t[t.Reaction_who == s].Reaction_what.value_counts().plot(title=s, kind='barh')
plt.figsize(20,2)
plt.show()
```



We can also look at the reactions for each candidate based on who the **transcript** says is speaking. The only difference that is obvious here is that the moderator is getting flack for some negative reactions for the candidates.

In [150]:
```
for s,n in enumerate(['Moderator','Romney','Obama']):
    plt.subplot(131+s)
    t[t.Speaker == s].Reaction_what.value_counts().plot(title=n, kind='barh')
plt.figsize(20,2)
```



Let's make a column that gets the speaker name instead of a number.

In [262]:
```
t['name'] = t.Speaker.apply(lambda s: {0:'Moderator', 1:'Romney', 2:'Obama'}[s])
t[['statement','Speaker','name']].groupby('Speaker').head(2)
```

Out[262]:

		statement	Speaker	name
Speaker				
0	**0**	0	0	Moderator
	56861	0	0	Moderator
1	**43**	43	1	Romney

	44	44	1	Romney
2	**20**	20	2	Obama
	141610	20	2	Obama

Text and reaction counts for each turn

Get all the transcript text for each turn as one string, and the number of reactions for that turn. First get the first transcript for each statement (they are all the same).

In [263]:
```
t2 = t.groupby(['statement']).first()[['name','Transcript','turn']]
t2.head(2)
```

Out[263]:

	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

A new dataframe with a row for each turn having the full turn's transcript and other statistics.

In [266]:
```
t3 = pd.DataFrame({'name':t2.groupby('turn').first().name, 'reactions':t.groupby('turn').count().Time, 'statements':t2.groupby('turn').count
t3['r_per_s'] = 1.0 * t3.reactions / t3.statements
t3.head(2)
```

Out[266]:

	name	reactions	statements	text	r_per_s
turn					
1	Moderator	812	20	Good evening from the Magness Arena at the Uni...	40.6
2	Obama	4213	22	Well, thank you very much, Jim, for this oppor...	191.5

Text features

Get a clean list of words for each turn and get a ranked list of the most common words so that we can decide which words will be used as features. We want to control how many can become features to be able to avoid overfitting.

In [308]:
```
t3['words'] = t3.text.apply(lambda txt: [t.lower() for t in nltk.tokenize.word_tokenize(txt) if t.isalpha()])
t3['word_count'] = t3.words.apply(lambda words: len(words))
ranked_unigrams = nltk.FreqDist([w for word_list in t3.words for w in word_list]).keys()
ranked_unigrams[:10]
```

Out[308]:
```
['the', 'to', 'that', 'and', 'of', 'a', 'we', 'i', 'you', 'in']
```

Make a unigrams features list for each turn.

In [309]:
```
MAX_FEATURES = 1000
t3['unigrams'] = t3.words.apply(lambda words: {w:True for w in words if w in ranked_unigrams[:MAX_FEATURES]})
t3['unigram_count'] = t3.unigrams.apply(lambda unigrams: len(unigrams))
t3.head(2)
```

Out[309]:

	name	reactions	statements	text	r_per_s	words	word_count	unigrams	unigram_count	label
turn										
1	Moderator	812	20	Good evening from the Magness Arena at the Uni...	40.6	[good, evening, from, the, magness, arena, at,...	257	{'all': True, 'domestic': True, 'questions': T...	115	False
2	Obama	4213	22	Well, thank you very much, Jim, for this oppor...	191.5	[well, thank, you, very, much, jim, for, this,...	278	{'sector': True, 'all': True, 'code': True, 'j...	137	True

Looking at features and cleanup

In [310]:
```
t3.describe()
```

Out[310]:

	reactions	statements	r_per_s	word_count	unigram_count
count	190.000000	190.000000	190.000000	190.000000	190.000000
mean	994.815789	6.168421	111.787644	77.452632	39.115789
std	1622.915791	8.907107	81.179389	121.482315	52.353168
min	1.000000	1.000000	1.000000	0.000000	0.000000
25%	53.500000	1.000000	43.125000	5.000000	4.000000

25%	53.500000	1.000000	43.125000	3.000000	4.000000
50%	141.500000	2.000000	92.000000	14.000000	11.500000
75%	1075.500000	6.750000	167.028846	84.000000	54.500000
max	7777.000000	54.000000	393.500000	497.000000	195.000000

There are turns with zero words? I wonder what those turns are? It turns out there is only one. I'm not sure why tokenization didn't split up this sentence. Anyway, let's get rid of it.

```
In [311]: print len(t3[t3.word_count == 0])
          print t3[t3.word_count == 0].text.values

          1
          [Absolutely.Yes.Absolutely.]
```

This still leaves some turns that are pretty laconic, but those might be good ones..

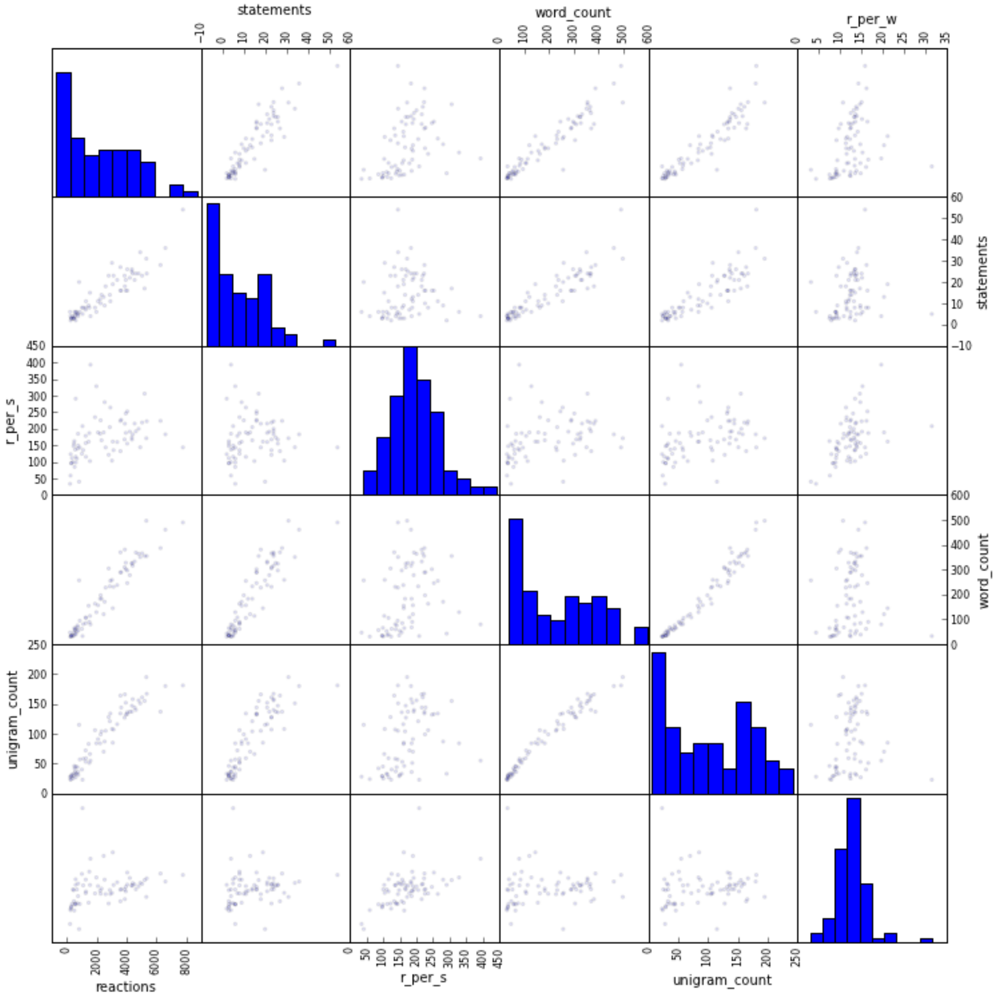
```
In [320]: #MIN_WORDS = 10
          MIN_WORDS = 30
          t4 = t3[t3.word_count >= MIN_WORDS]
          print len(t3),'->',len(t4)

          190 -> 71
```

Let's look at reactions per word as well.

```
In [321]: t4['r_per_w'] = 1.0 * t4.reactions / t4.word_count
```

Let's see what features and stats we now have.

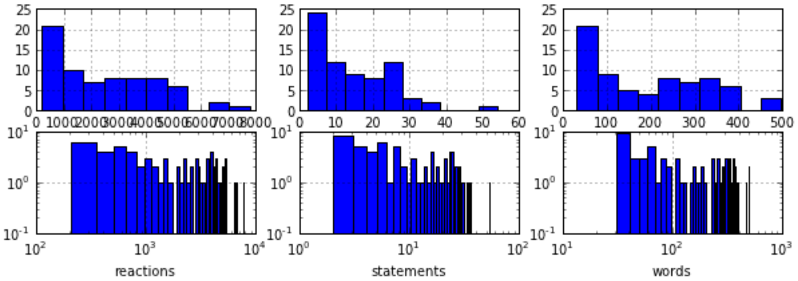
```
In [322]: scatter_matrix(t4, alpha=1.0/10, figsize=(12, 12), diagonal='hist')
          show()
```



View a few with a log transform.

```
In [323]: subplot(231)
          t4.reactions.hist()
          xlabel('reactions')

          subplot(234)
          t4.reactions.hist(bins=50, log=True)
          xscale('log')
          xlabel('reactions')
```

```
subplot(232)
t4.statements.hist()
xlabel('statements')

subplot(235)
t4.statements.hist(bins=50, log=True)
xscale('log')
xlabel('statements')

subplot(233)
t4.word_count.hist()
xlabel('words')

subplot(236)
t4.word_count.hist(bins=50, log=True)
xscale('log')
xlabel('words')

figsize(15,6)
```



The number of words per turn is interesting. There seems to be a large number of turns with just 10 or so words, then a few with another bump at ~350.

What to predict?

Bin the turns into categories; we'll later try to predict the category based on the text features.

Reactions per statement

One option is by reactions per statement. What are examples of turns with high reactions per statement and low r/s?

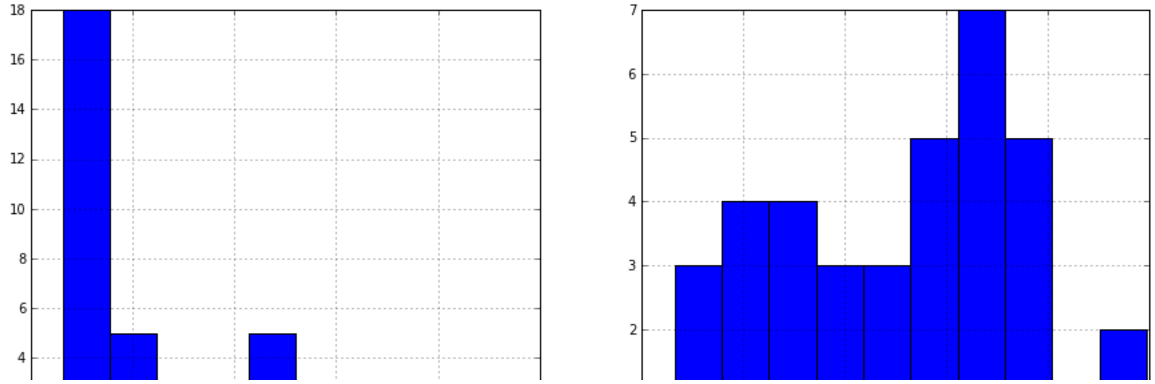
In [325]:
```
PERC = .10
print '{:_^80}'.format('high R/S'.upper())
for v in t4[t4.r_per_s > t4.r_per_s.quantile(1-PERC)].text.values: print v+'\n'
print '{:_^80}'.format('low R/S'.upper())
for v in t4[t4.r_per_s < t4.r_per_s.quantile(PERC)].text.values: print v+'\n'
```

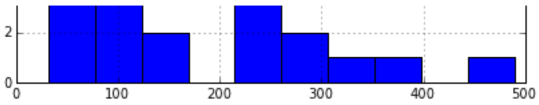
```
____________________________________HIGH R/S____________________________________
When you add up all the loopholes and deductions that upper income individuals are currently taking advantage ofif you take those all away,
you don't come close to paying for $5 trillion in tax cuts and $2 trillion in additional military spending.And that's why independent
studies looking at this said the only way to meet Governor Romney's pledge of not reducing the deficit or not adding to the deficit, is by
burdening middle-class families.The average middle-class family with children would pay about $2,000 more.Now, that's not my analysis;
that's the analysis of economists who have looked at this.And that kind of top-down economics, where folks at the top are doing well so the
average person making 3 million bucks is getting a $250,000 tax break while middle-class families are burdened further, that's not what I
believe is a recipe for economic growth.

Jim, I -- you may want to move on to another topic, but I would just say this to the American people.If you believe that we can cut taxes
by $5 trillion and add $2 trillion in additional spending that the military is not asking for -- $7 trillion, just to give you a sense,
over 10 years that's more than our entire defense budgetand you think that by closing loopholes and deductions for the well-to-do, somehow
you will not end up picking up the tab, then Governor Romney's plan may work for you.But I think math, common sense and our history shows
us that's not a recipe for job growth.Look, we've tried this -- we've tried both approaches.The approach that Governor Romney's talking
about is the same sales pitch that was made in 2001 and 2003.And we ended up with the slowest job growth in 50 years.We ended up moving
from surplus to deficits.And it all culminated in the worst financial crisis since the Great Depression.Bill Clinton tried the approach
that I'm talking about.We created 23 million new jobs.We went from deficit to surplus, and businesses did very well.So in some ways, we've
got some data on which approach is more likely to create jobs and opportunity for Americansand I believe that the economy works best when
middle-class families are getting tax breaks so that they've got some money in their pocketsand those of us who have done extraordinarily
well because of this magnificent country that we live in, that we can afford to do a little bit more to make sure we're not blowing up the
deficit.

It means that -- Governor Romney talked about Medicaid and how we could send it back to the statesbut effectively this means a 30 percent
cut in the primary program we help for seniors who are in nursing homes, for kids who are with disabilities --

No, I -- I think I've -- I had five seconds before you interrupted me. That the irony is that we've seen this model work really well in
Massachusetts, because Governor Romney did a good thing, working with Democrats in the state to set up what is essentially the identical
model.And as a consequence, people are covered there, it hasn't destroyed jobs.And as a consequence, we now have a system in which we have
the opportunity to start bringing down cost, as opposed to just --

The -- where their partnering so that -- they're designing training programs, and people who are going through them know that there's a job
waiting for them if they complete them. That makes a big difference but that requires some federal support.Let me just say one final
example. When it comes to making college affordable -- whether it's two-year or four-year -- one of the things that I did as president was
we were sending $60 billion to banks and lenders as middle men for the student loan program, even though the loans were guaranteed. So
there was no risk for the banks or the lenders but they were taking billions out of the system.And we said, why not cut out the middle man?
And as a consequence, what we've been able to do is to provide millions more students assistance, lower or keep low interest rates on
student loans. And this is an example of where our priorities make a difference. Governor Romney, I genuinely believe, cares about
```

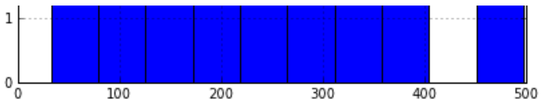
```
education. But when he tells a student that, you know, you should borrow money from your parents to go to college, you know, that indicates
the degree to which, you know, there may not be as much of a focus on the fact that folks like myself, folks like Michelle, kids probably
who attend University of Denver just don't have that option.And for us to be able to make sure that they've got that opportunity and they
can walk through that door, that is vitally important -- not just to those kids. It's how we're going to grow this economy over the long
term.

We -- as president, I will sit down on day one -- actually the day after I get elected, I'll sit down with leaders -- the Democratic
leaders as well as Republican leaders and -- as we did in my state. We met every Monday for a couple hours, talked about the issues and the
challenges in our state, in that case. We have to work on a collaborative basis -- not because we're going to compromise our principle(s),
but because there's common ground.And the challenges America faces right now -- look, the reason I'm in this race is there are people that
are really hurting today in this country, and we face -- this deficit could crush the future generations. What's happening in the Middle
East? There are developments around the world that are of real concern. And Republicans and Democrats both love America, but we need to
have leadership in Washington that will actually bring people together and get the job done and could not care less if it's a Republican or
a Democrat. I've done it before. I'll do it again.

Well, first of all, I think Governor Romney's going to have a busy first day, because he's also going to repeal �Obamacare,� which will
not be very popular among Democrats as you're sitting down with them.But look, my philosophy has been I will take ideas from anybody,
Democrat or Republican, as long as they're advancing the cause of making middle-class families stronger and giving ladders of opportunity
into the middle class. That's how we cut taxes for middle-class families and small businesses. That's how we cut a trillion dollars of
spending that wasn't advancing that cause. That's how we signed three trade deals into law that are helping us to double our exports and
sell more American products around the world. That's how we repealed �don't ask, don't tell.� That's how we ended the war in Iraq, as I
promised and that's how we're going to wind down the war in Afghanistan. That's how we went after al-Qaida and bin Laden.So we've seen
progress even under Republican control of the House or Representatives. But ultimately, part of being principled, part of being a leader
is, A, being able to describe exactly what it is that you intend to do, not just saying, I'll sit down, but you have to have a plan.Number
two, what's important is occasionally you've got to say no to folks both in your own party and in the other party. And you know, yes, have
we had some fights between me and the Republicans when they fought back against us, reining in the excesses of Wall Street? Absolutely,
because that was a fight that needed to be had. When we were fighting about whether or not we were going to make sure that Americans had
more security with their health insurance and they said no, yes, that was a fight that we needed to have. And so part of leadership and
governing is both saying what it is that you are for, but also being willing to say no to some things.And I've got to tell you, Governor
Romney, when it comes to his own party during the course of this campaign, has not displayed that willingness to say no to some of the more
extreme parts of his party.

___________________________________LOW R/S____________________________________
Good evening from the Magness Arena at the University of Denver in Denver, Colorado.I'm Jim Lehrer of the PBS NewsHour, and I welcome you
to the first of the 2012 presidential debates between President Barack Obama, the Democratic nominee, and former Massachusetts Governor
Mitt Romney, the Republican nominee.This debate and the next three -- two presidential, one vice- presidential -- are sponsored by the
Commission on Presidential Debates.Tonight's 90 minutes will be about domestic issues, and will follow a format designed by the
commission.There will be six roughly 15-minute segments, with two-minute answers for the first question, then open discussion for the
remainder of each segment.Thousands of people offered suggestions on segment subjects or questions via the Internet and other means but I
made the final selectionsAnd for the record, they were not submitted for approval to the commission or the candidates.The segments, as I
announced in advance, will be three on the economy and one each on health care, the role of government, and governing, with an emphasis
throughout on differences, specifics and choices.Both candidates will also have two-minute closing statements.The audience here in the hall
has promised to remain silent.No cheers, applause, boos, hisses -- among other noisy distracting things -- so we may all concentrate on
what the candidates have to say.There is a noise exception right now, though, as we welcome President Obama and Governor Romney. Gentlemen,
welcome to you both.Let's start the economy, segment one.And let's begin with jobs.What are the major differences between the two of you
about how you would go about creating new jobs?You have two minutes -- each of you have two minutes to start.The coin toss has determined,
Mr. President, you go first.

It's OK. It's great.That's OK.No problem.No, you don't have -- you don't have a problem, I don't have a problem, because we're still on the
economybut we're going to come back to taxes and we're going to move on to the deficit and a lot of other things, too.OK, but go ahead,
sir.

First of all, the Department of Energy has said the tax break for oil companies is $2.8 billion a year.And it's actually an accounting
treatment, as you know, that's been in place for a hundred years. Now --

All right? All right, this is this is segment three, the economy, entitlements.First answer goes to you.It's two minutes.Mr. President, do
you see a major difference between the two of you on Social Security?

on Medicare?All right.So, to finish quickly, briefly, on the economy, what is your view about the level of federal regulation of the
economy right now?Is there too much, and in your case, Mr. President, is there -- should there be more?Beginning with you -- this is not a
new two-minute segment -- to start, and we'll go for a few minutes and then we're going to go to health care.OK?

All right, I think we have another clear difference between the two of you.Now let's move to health care, where I know there is a clear
difference -- and that has to do with the Affordable Care Act, �Obamacare.�And it's a two-minute new segment, and it's -- that means two
minutes each.And you go first, Governor Romney.You wanted repeal.You want the Affordable Care Act repealed. Why?

Excuse me, one sec -- excuse, me sir. We've got barely have three minutes left. I'm not going to grade the two of you and say your answers
have been too long or I've done a poor job --
```

This is a bad idea it seems because there are some turns only have really short statements. Clearly, some of those should be turns, either..

```
In [326]: subplot(121)
          t4[t4.r_per_s < t4.r_per_s.quantile(.5)].word_count.hist()
          subplot(122)
          t4[t4.r_per_s >= t4.r_per_s.quantile(.5)].word_count.hist()
          figsize(10,3)
```







Reactions per word

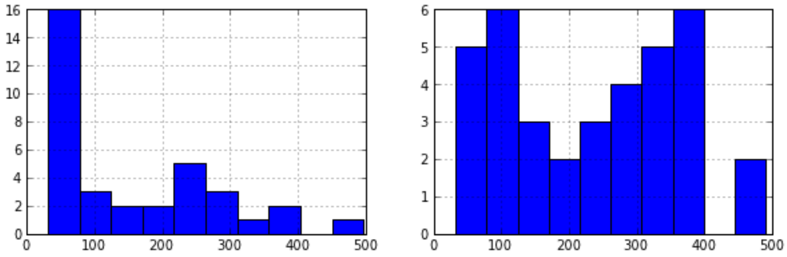
Another option is reactions per word for each turn.

```
In [327]: PERC = .10
          print '{:_^80}'.format('high r/w'.upper())
          for v in t4[t4.r_per_w > t4.r_per_w.quantile(1-PERC)].text.values: print v+'\n'
          print '{:_^80}'.format('low r/w'.upper())
          for v in t4[t4.r_per_w < t4.r_per_w.quantile(PERC)].text.values: print v+'\n'
```

```
____________________________________HIGH R/W____________________________________
Jim, I -- you may want to move on to another topic, but I would just say this to the American people.If you believe that we can cut taxes
by $5 trillion and add $2 trillion in additional spending that the military is not asking for -- $7 trillion, just to give you a sense,
over 10 years that's more than our entire defense budgetand you think that by closing loopholes and deductions for the well-to-do, somehow
you will not end up picking up the tab, then Governor Romney's plan may work for you.But I think math, common sense and our history shows
us that's not a recipe for job growth.Look, we've tried this -- we've tried both approaches.The approach that Governor Romney's talking
about is the same sales pitch that was made in 2001 and 2003.And we ended up with the slowest job growth in 50 years.We ended up moving
from surplus to deficits.And it all culminated in the worst financial crisis since the Great Depression.Bill Clinton tried the approach
that I'm talking about.We created 23 million new jobs.We went from deficit to surplus, and businesses did very well.So in some ways, we've
got some data on which approach is more likely to create jobs and opportunity for Americansand I believe that the economy works best when
middle-class families are getting tax breaks so that they've got some money in their pocketsand those of us who have done extraordinarily
well because of this magnificent country that we live in, that we can afford to do a little bit more to make sure we're not blowing up the
deficit.

Good.OK, good.So I'll get rid of that.I'm sorry, Jim.I'm going to stop the subsidy to PBS.I'm going to stop other things.I like PBS.I like
Big Bird.I actually like you too.But I'm not going to -- I'm not going to keep on spending money on things to borrow money from China to
pay for it.That's number one.Number two, I'll take programs that are currently good programs but I think could be run more efficiently at
the state level and send them to state.Number three, I'll make government more efficient, and to cut back the number of employees, combine
some agencies and departments.My cutbacks will be done through attrition, by the way.This is the approach we have to take to get America to
a balanced budget.The president said he'd cut the deficit in half.Unfortunately, he doubled it.Trillion-dollar deficits for the last four
years.The president's put it in place as much public debt -- almost as much debt held by by the public as all prior presidents combined.

I -- look, the revenue I get is by more people working, getting higher pay, paying more taxes.That's how we get growth and how we balance
the budget.But the idea of taxing people more, putting more people out of work -- you'll never get there.You never balance the budget by
raising taxes.Spain -- Spain spends 42 percent of their total economy on government.We're now spending 42 percent of our economy on
government.I don't want to go down the path to Spain.I want to go down the path of growth that puts Americans to work, with more money
coming in because the're working.

No, I -- I think I've -- I had five seconds before you interrupted me. That the irony is that we've seen this model work really well in
Massachusetts, because Governor Romney did a good thing, working with Democrats in the state to set up what is essentially the identical
model.And as a consequence, people are covered there, it hasn't destroyed jobs.And as a consequence, we now have a system in which we have
the opportunity to start bringing down cost, as opposed to just --

Well, first, I love great schools. Massachusetts, our schools are ranked number one of all 50 states. And the key to great schools: great
teachers. So I reject the idea that I don't believe in great teachers or more teachers. Every school district, every state should make that
decision on their own.The role of government -- look behind us: the Constitution and the Declaration of Independence.The role of government
is to promote and protect the principles of those documents. First, life and liberty. We have a responsibility to protect the lives and
liberties of our people, and that means the military, second to none. I do not believe in cutting our military. I believe in maintaining
the strength of America's military.Second, in that line that says, we are endowed by our Creator with our rights -- I believe we must
maintain our commitment to religious tolerance and freedom in this country. That statement also says that we are endowed by our Creator
with the right to pursue happiness as we choose. I interpret that as, one, making sure that those people who are less fortunate and can't
care for themselves are cared by one another.We're a nation that believes we're all children of the same God. Look, the genius of America
is the free enterprise system, and freedom, and the fact that people can go out there and start a business, work on an idea, make their own
decisions.And we care for those that have difficulties -- those that are elderly and have problems and challenges, those that disabled, we
care for them. And we look for discovery and innovation, all these thing desired out of the American heart to provide the pursuit of
happiness for our citizens.But we also believe in maintaining for individuals the right to pursue their dreams, and not to have the
government substitute itself for the rights of free individuals. And what we're seeing right now is, in my view, a trickle-down government
approach which has government thinking it can do a better job than free people pursuing their dreams. And it's not working.And the proof of
that is 23 million people out of work. The proof of that is one out of six people in poverty. The proof of that is we've gone from 32
million on food stamps to 47 million on food stamps. The proof of that is that 50 percent of college graduates this year can't find work.We
know that the path we're taking is not working. It's time for a new path.

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.

Thank you, Governor.Thank you, Mr. President.The next debate will be the vice presidential event on Thursday, October 11th at Center
College in Danville, Kentucky. For now, from the University of Denver, I'm Jim Lehrer. Thank you, and good night.

____________________________________LOW R/W_____________________________________
Good evening from the Magness Arena at the University of Denver in Denver, Colorado.I'm Jim Lehrer of the PBS NewsHour, and I welcome you
to the first of the 2012 presidential debates between President Barack Obama, the Democratic nominee, and former Massachusetts Governor
Mitt Romney, the Republican nominee.This debate and the next three -- two presidential, one vice- presidential -- are sponsored by the
Commission on Presidential Debates.Tonight's 90 minutes will be about domestic issues, and will follow a format designed by the
commission.There will be six roughly 15-minute segments, with two-minute answers for the first question, then open discussion for the
remainder of each segment.Thousands of people offered suggestions on segment subjects or questions via the Internet and other means but I
made the final selectionsAnd for the record, they were not submitted for approval to the commission or the candidates.The segments, as I
announced in advance, will be three on the economy and one each on health care, the role of government, and governing, with an emphasis
throughout on differences, specifics and choices.Both candidates will also have two-minute closing statements.The audience here in the hall
has promised to remain silent.No cheers, applause, boos, hisses -- among other noisy distracting things -- so we may all concentrate on
what the candidates have to say.There is a noise exception right now, though, as we welcome President Obama and Governor Romney. Gentlemen,
welcome to you both.Let's start the economy, segment one.And let's begin with jobs.What are the major differences between the two of you
about how you would go about creating new jobs?You have two minutes -- each of you have two minutes to start.The coin toss has determined,
Mr. President, you go first.
```

```
It's OK. It's great.That's OK.No problem.No, you don't have -- you don't have a problem, I don't have a problem, because we're still on the
economybut we're going to come back to taxes and we're going to move on to the deficit and a lot of other things, too.OK, but go ahead,
sir.

Jim, let's -- we -- we've gone on a lot of topics there, and -- so I've got to take -- it's going to take a minute to go from Medicaid to
schools to --(Inaudible.)

First of all, the Department of Energy has said the tax break for oil companies is $2.8 billion a year.And it's actually an accounting
treatment, as you know, that's been in place for a hundred years. Now --

All right? All right, this is this is segment three, the economy, entitlements.First answer goes to you.It's two minutes.Mr. President, do
you see a major difference between the two of you on Social Security?

Excuse me, one sec -- excuse, me sir. We've got barely have three minutes left. I'm not going to grade the two of you and say your answers
have been too long or I've done a poor job --

Jim, I had the great experience -- it didn't seem like it at the time -- of being elected in a state where my legislature was 87 percent
Democrat, and that meant I figured out from day one I had to get along and I had to work across the aisle to get anything done. We drove
our schools to be number one in the nation. We cut taxes 19 times.
```

In [328]:
```
subplot(121)
t4[t4.r_per_w < t4.r_per_w.quantile(.5)].word_count.hist()
subplot(122)
t4[t4.r_per_w >= t4.r_per_w.quantile(.5)].word_count.hist()
figsize(10,3)
```



The low R/W category still seems to be dominated by turns with very short words.

In [329]:
```
#t43['label'] = t4.reactions >= t4.reactions.quantile(.5)
#t4['label'] = t4.r_per_s >= t4.r_per_s.quantile(.5)
t4['label'] = t4.r_per_w >= t4.r_per_w.quantile(.5)
print t4.r_per_w.quantile(.5)
t4.label.describe()
```

```
12.5176151762
```

Out[329]:
```
count           71
mean     0.5070423
std      0.5035088
min          False
25%              0
50%              1
75%              1
max           True
```

Train and test

In [330]:
```
train_rows = random.sample(t4.index, len(t4)*9/10)
trn = t4.ix[train_rows]
tst = t4.drop(train_rows)
print len(trn)
print len(tst)
```

```
63
8
```

In [331]:
```
%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: 0.07 s
```

In [332]:
```
nltk.classify.accuracy(cl, zip(tst.unigrams, tst.label))
```

Out[332]:
```
0.625
```

In [333]:
```
cl.show_most_informative_features()
```

```
Most Informative Features
                 million = True              True : False  =      7.4 : 1.0
                approach = True              True : False  =      5.5 : 1.0
                families = True              True : False  =      4.8 : 1.0
                   means = True              True : False  =      4.2 : 1.0
                      in = None             False : True   =      3.9 : 1.0
                      we = None             False : True   =      3.9 : 1.0
                    last = True              True : False  =      3.7 : 1.0
                american = True              True : False  =      3.7 : 1.0
                 federal = True             False : True   =      3.7 : 1.0
```

```
                              most = True              True : False   =      3.6 : 1.0
```

This is really garbage. I don't see any meaning in these features.

Multiple train/test partitions

In [344]:

```
accs = []
for i in range(10):
    train_rows2 = random.sample(t4.index, len(t4)*9/10)
    trn2 = t4.ix[train_rows2]
    tst2 = t4.drop(train_rows2)
    cl2 = nltk.NaiveBayesClassifier.train(zip(trn2.unigrams, trn2.label))
    accs.append(nltk.classify.accuracy(cl2, zip(tst2.unigrams, tst2.label)))
a2 = pd.DataFrame({'accuracy':accs})
print a2
```

```
   accuracy
0     0.375
1     0.625
2     0.250
3     0.375
4     0.625
5     0.625
6     0.250
7     0.500
8     0.500
9     0.625
```

In [345]:

```
print a2.describe()
```

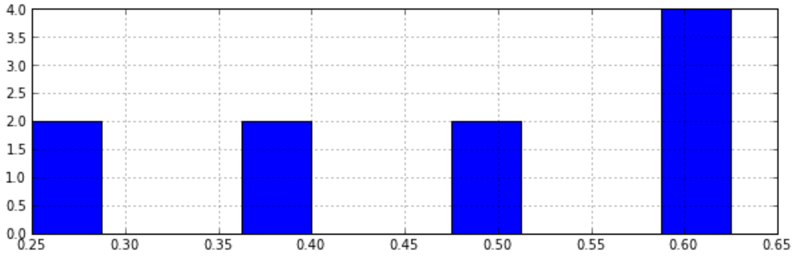
```
        accuracy
count  10.000000
mean    0.475000
std     0.153659
min     0.250000
25%     0.375000
50%     0.500000
75%     0.625000
max     0.625000
```

In [346]:

```
a2.accuracy.hist()
```

Out[346]:

```
<matplotlib.axes.AxesSubplot at 0x9402a30>
```



Feature hyperparams

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

In [347]:

```
t5 = t4.copy()
```

In [348]:

```
len(ranked_unigrams)
```

Out[348]:

```
1757
```

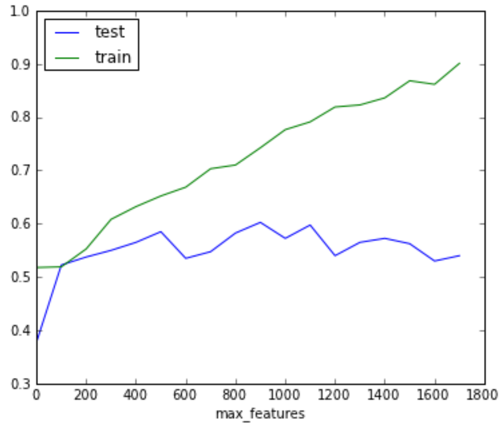
In [349]:

```
p = []
trn_means = []
tst_means = []
for max_feats in range(1,len(ranked_unigrams),100):
    t5['unigrams'] = t5.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_rows2 = random.sample(t5.index, len(t5)*9/10)
        trn2,tst2 = t5.ix[train_rows2],t5.drop(train_rows2)
        cl2 = nltk.NaiveBayesClassifier.train(zip(trn2.unigrams, trn2.label))
        trn_ac.append(nltk.classify.accuracy(cl2, zip(trn2.unigrams, trn2.label)))
        tst_ac.append(nltk.classify.accuracy(cl2, zip(tst2.unigrams, tst2.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 [359]:
```
figsize(6,5)
results = pd.DataFrame({'max_features':p, 'train':trn_means, 'test':tst_means})
results.plot(x='max_features')
```

Out[359]:
```
<matplotlib.axes.AxesSubplot at 0x734c850>
```



In [360]:
```
results
```

Out[360]:

	max_features	test	train
0	1	0.3775	0.517778
1	101	0.5225	0.519048
2	201	0.5375	0.552698
3	301	0.5500	0.608254
4	401	0.5650	0.632063
5	501	0.5850	0.652063
6	601	0.5350	0.668571
7	701	0.5475	0.703175
8	801	0.5825	0.710159
9	901	0.6025	0.742540
10	1001	0.5725	0.776508
11	1101	0.5975	0.791111
12	1201	0.5400	0.819048
13	1301	0.5650	0.822857
14	1401	0.5725	0.836190
15	1501	0.5625	0.868254
16	1601	0.5300	0.861587

16	1601	0.5300	0.861587
17	1701	0.5400	0.900952

It looks like going past ~900 unigram features is not helpful.

For now, this is the best we have as a unigram baseline for this task. Improving on this might take some other creative idea, or changing the classifier model.

In [ ]: