Department of Computer Science
UNIVERSITY OF COLORADO **BOULDER**

Classification: Feature Engineering

Machine Learning: Jordan Boyd-Graber
University of Colorado Boulder
LECTURE 4

Content Questions

Content Questions

Content Questions

Content Questions

Content Questions

Content Questions

Content Questions

Content Questions

Content Questions

Content Questions

Announcements

- HW1 Turned in
- HW2 Due on Friday
- HW3 Due next week

Administrivia Questions

Administrivia Questions

Administrivia Questions

Feature Engineering

- Questions from trivia games
- Divided into specific categories
- Important to know which is which

Words

Listing 1 : Words

```
History:battles:efforts:pyramid:russia:byzantine:presidency:organized:foreign
Literature:poetic:jew:poet:poem:novel:poems:stories:literary:playwright:novels
Social Science:holiday:ritual:zeus:psychological:economist:philosopher:
     anthropologist:rivers:psychologist:deity
Geography:built:nile:hill:square:fault:feature:mountain:2011:bridge:red
Other:code:win:movies:strip:stanley:season:starring:oscar:yr:rl
Biology:selection:enzyme:humans:chromosome:membrane:amino:syndrome:cellular:plants
Fine Arts:painter:canvas:composers:foreground:commissioned:piano:painting:opera
Physics:physical:vector:scattering:classical:quark:physics:parameter:physicist
Chemistry:paradox:doubly:obtained:concentrations:molar:spectroscopy:ion:presence
Mathematics:denoted:curve:algebraic:differential:polygon:dimensional:euler:methods
Earth Science:areas:hot:discontinuity:earth:region:period:rock:material:mantle
Science:momentum:largest:contrasted:radiation:limit:magnitude:origin:increase:equals
Astronomy:found:beyond:object:astronomer:constellation:phenomenon:stars:objects
```

Words

Listing 2 : Words

```
History:battles:efforts:pyramid:russia:byzantine:presidency:organized:foreign
Literature:poetic:jew:poet:poem:novel:poems:stories:literary:playwright:novels
Social Science:holiday:ritual:zeus:psychological:economist:philosopher:
     anthropologist:rivers:psychologist:deity
Geography:built:nile:hill:square:fault:feature:mountain:2011:bridge:red
Other:code:win:movies:strip:stanley:season:starring:oscar:yr:rl
Biology:selection:enzyme:humans:chromosome:membrane:amino:syndrome:cellular:plants
Fine Arts:painter:canvas:composers:foreground:commissioned:piano:painting:opera
Physics:physical:vector:scattering:classical:quark:physics:parameter:physicist
Chemistry:paradox:doubly:obtained:concentrations:molar:spectroscopy:ion:presence
Mathematics:denoted:curve:algebraic:differential:polygon:dimensional:euler:methods
Earth Science:areas:hot:discontinuity:earth:region:period:rock:material:mantle
Science:momentum:largest:contrasted:radiation:limit:magnitude:origin:increase:equals
Astronomy:found:beyond:object:astronomer:constellation:phenomenon:stars:objects
```

Accuracy: 0.765

Bigrams Only

Listing 3 : Bigrams Only

```
History:who served:vice presidential:the election:chief of:civil rights:b c
Literature:his novels:this literary:his plays:poem about:the novel:this poem
Social Science:this economic:this anthropologist:of social:cities include
Geography:this state:lake s:the nile:the oldest:was built:country that:this lake
Other:in 2005:this show:an oscar:comic strip:yr american:the album:played by
Biology:amino acid:syndrome and:the enzyme:the brain:on chromosome:of cell
Fine Arts:this composer:painter of:opera about:its composer:this painting:this opera
Physics:formula for:a quantum:x rays:of matter:of particles:the universe:to measure
Chemistry:chemist who:is present:this chemist:tend to:the molar:a carbonyl
Mathematics:sub n:data structure:algorithm for:mathematician whose:proof of
Earth Science:its name:mohs hardness:above it:characterized by:the paleozoic
Science:this process:this number:reaction this:stored in:the equation:limit of
Astronomy:the solar:the universe:moon of:the milky:milky way:of objects:solar system
```

Bigrams Only

Listing 4 : Bigrams Only

```
History:who served:vice presidential:the election:chief of:civil rights:b c
Literature:his novels:this literary:his plays:poem about:the novel:this poem
Social Science:this economic:this anthropologist:of social:cities include
Geography:this state:lake s:the nile:the oldest:was built:country that:this lake
Other:in 2005:this show:an oscar:comic strip:yr american:the album:played by
Biology:amino acid:syndrome and:the enzyme:the brain:on chromosome:of cell
Fine Arts:this composer:painter of:opera about:its composer:this painting:this opera
Physics:formula for:a quantum:x rays:of matter:of particles:the universe:to measure
Chemistry:chemist who:is present:this chemist:tend to:the molar:a carbonyl
Mathematics:sub n:data structure:algorithm for:mathematician whose:proof of
Earth Science:its name:mohs hardness:above it:characterized by:the paleozoic
Science:this process:this number:reaction this:stored in:the equation:limit of
Astronomy:the solar:the universe:moon of:the milky:milky way:of objects:solar system
```

Accuracy: 0.800

Trigrams Only

Listing 5 : Trigrams Only

```
History:the battle of:name this leader:ftp what was:an alliance with:ruler of this
Literature:identify this author:this short story:of the play:t s eliot:in this play
Social Science:this deity s:name this economic:this thinker s:of this concept:this
     son of:name this concept:name this greek:the holy spirit:the concept of
Geography:name this largest:this state contains:this island s:this river s:this
     geological feature:of this mountain:of this river:this state s:this lake s
Other:on the album:name this film:this comic strip:hall of fame:name this band:name
     this actress:of the year:in this film:character played by:in the film
Biology:disease of the:name this phylum:the production of:they have a:this amino
     acid:in e coli:of the brain:is responsible for:name this enzyme
Fine Arts:of this painting:this composer of:of this opera:name this composer
Physics:is governed by:the standard model:name this physicist:prize in physics
Chemistry:this doubly eponymous:the presence of:of organic compounds
Mathematics:name these mathematical:product of two:name this topological:the set of
Earth Science:of the paleozoic:d double prime:of this mineral:name this mineral:of
     this period:the earth s:of this rock:of the earth:mohs hardness of:million
     years ago
Science:is broken down:this compound s:the photoelectric effect:these organelles are
     :ones have a:to this quantity:these objects is:this is the:is measured in:this
      element is
Astronomy:a black hole:in the sky:of the universe:brightest star is:in the
     constellation:largest moon of:the main sequence:from the sun:the milky way:the
      solar system
```

Trigrams Only

Listing 6 : Trigrams Only

```
History:the battle of:name this leader:ftp what was:an alliance with:ruler of this
Literature:identify this author:this short story:of the play:t s eliot:in this play
Social Science:this deity s:name this economic:this thinker s:of this concept:this
     son of:name this concept:name this greek:the holy spirit:the concept of
Geography:name this largest:this state contains:this island s:this river s:this
     geological feature:of this mountain:of this river:this state s:this lake s
Other:on the album:name this film:this comic strip:hall of fame:name this band:name
     this actress:of the year:in this film:character played by:in the film
Biology:disease of the:name this phylum:the production of:they have a:this amino
     acid:in e coli:of the brain:is responsible for:name this enzyme
Fine Arts:of this painting:this composer of:of this opera:name this composer
Physics:is governed by:the standard model:name this physicist:prize in physics
Chemistry:this doubly eponymous:the presence of:of organic compounds
Mathematics:name these mathematical:product of two:name this topological:the set of
Earth Science:of the paleozoic:d double prime:of this mineral:name this mineral:of
     this period:the earth s:of this rock:of the earth:mohs hardness of:million
     years ago
Science:is broken down:this compound s:the photoelectric effect:these organelles are
     :ones have a:to this quantity:these objects is:this is the:is measured in:this
      element is
Astronomy:a black hole:in the sky:of the universe:brightest star is:in the
     constellation:largest moon of:the main sequence:from the sun:the milky way:the
      solar system
```

Accuracy: 0.756

Unigrams and Bigrams

Listing 7 : Unigrams and Bigrams

```
History:khan:emperor:successor:occurred:minister:byzantine:capture:empire:dynasty
Literature:play:poet:playwright:stories:author who:literature:novel:10 points
Social Science:bull:economist:holiday:anthropologist:psychological:psychologist:
     philosopher:rivers:ritual:deity
Geography:african:was built:bridge:red:sea:city s:this lake:mountain:feature
Other:band:player:season:film:starring:team:oscar:movie:yr:rl
Biology:cellular:membrane:protein:cell:syndrome:chromosome:enzyme:dna:plants:genes
Fine Arts:composition:movement:painting:piano:painted:opera:10 pointsname:pointsname
Physics:matter:voltage:quark:physical:wavelength:physicist:materials:physics:spin
Chemistry:chemical:reacts:spectroscopy:ion:gases:molar:reaction:this equation
Mathematics:euler:matrix:curve:prime:mathematical:methods:problem:algebraic
Earth Science:zone:forms:period:hot:mantle:boundary:discontinuity:rock:mineral
Science:largest:this protein:this number:an electron:bond:magnitude:such as:angle
Astronomy:found:object:galaxies:mass:distance:atmosphere:star:astronomical:stars
```

Unigrams and Bigrams

Listing 8 : Unigrams and Bigrams

```
History:khan:emperor:successor:occurred:minister:byzantine:capture:empire:dynasty
Literature:play:poet:playwright:stories:author who:literature:novel:10 points
Social Science:bull:economist:holiday:anthropologist:psychological:psychologist:
     philosopher:rivers:ritual:deity
Geography:african:was built:bridge:red:sea:city s:this lake:mountain:feature
Other:band:player:season:film:starring:team:oscar:movie:yr:rl
Biology:cellular:membrane:protein:cell:syndrome:chromosome:enzyme:dna:plants:genes
Fine Arts:composition:movement:painting:piano:painted:opera:10 pointsname:pointsname
Physics:matter:voltage:quark:physical:wavelength:physicist:materials:physics:spin
Chemistry:chemical:reacts:spectroscopy:ion:gases:molar:reaction:this equation
Mathematics:euler:matrix:curve:prime:mathematical:methods:problem:algebraic
Earth Science:zone:forms:period:hot:mantle:boundary:discontinuity:rock:mineral
Science:largest:this protein:this number:an electron:bond:magnitude:such as:angle
Astronomy:found:object:galaxies:mass:distance:atmosphere:star:astronomical:stars
```

Accuracy: 0.803

Unigrams, Bigrams, Trigrams

Listing 9 : Unigrams

```
History:emperor:organized:occurred:russia:legislation:dynasty:treaty:successor:
     minister:empire
Literature:play:poet:writer:playwright:literature:poems:poem:novel:literary:novels
Social Science:hindu:anthropologist:holiday:demand:economist:ritual:psychologist:
     philosopher:rivers:deity
Geography:this river:of this river:was built:red:this lake:feature:bay:bridge:sea:
     mountain
Other:team:player:film:code:movie:season:yr:oscar:rl:starring
Biology:humans:syndrome:protein:cellular:gene:membrane:enzyme:dna:genes:plants
Fine Arts:10 pointsname this:pointsname this:pointsname this author:composition:
     painting:piano:movement:painted:opera:composer
Physics:wavelength:materials:decay:material:voltage:physics:spin:physicist:
     scattering:quark
Chemistry:hydrogen:molecules:this equation:spectroscopy:molar:ion:reaction:carbonyl:
     chemistry:chemist
Mathematics:named after:problem:space:data:algebraic:methods:curve:mathematical:
     mathematician:algorithm
Earth Science:zone:hot:boundary:areas:period:material:region:mineral:mantle:rock
Science:radiation:machine:momentum:this group:atom:square:velocity:bond:angle:equals
Astronomy:supernova:telescope:mass:astronomical:radiation:limit:galaxies:star:stars:
     objects
```

Unigrams, Bigrams, Trigrams

Listing 10 : Unigrams

```
History:emperor:organized:occurred:russia:legislation:dynasty:treaty:successor:
     minister:empire
Literature:play:poet:writer:playwright:literature:poems:poem:novel:literary:novels
Social Science:hindu:anthropologist:holiday:demand:economist:ritual:psychologist:
     philosopher:rivers:deity
Geography:this river:of this river:was built:red:this lake:feature:bay:bridge:sea:
     mountain
Other:team:player:film:code:movie:season:yr:oscar:rl:starring
Biology:humans:syndrome:protein:cellular:gene:membrane:enzyme:dna:genes:plants
Fine Arts:10 pointsname this:pointsname this:pointsname this author:composition:
     painting:piano:movement:painted:opera:composer
Physics:wavelength:materials:decay:material:voltage:physics:spin:physicist:
     scattering:quark
Chemistry:hydrogen:molecules:this equation:spectroscopy:molar:ion:reaction:carbonyl:
     chemistry:chemist
Mathematics:named after:problem:space:data:algebraic:methods:curve:mathematical:
     mathematician:algorithm
Earth Science:zone:hot:boundary:areas:period:material:region:mineral:mantle:rock
Science:radiation:machine:momentum:this group:atom:square:velocity:bond:angle:equals
Astronomy:supernova:telescope:mass:astronomical:radiation:limit:galaxies:star:stars:
     objects
```

Accuracy: 0.809

Character *n*-grams

Listing 11 : Character *n*-grams

```
History: khan: case: post: colo: new :orth : died:ses. : i : (+)
Literature: epic:novel: " : play :play : play: poem :poem : poet: poem
Social Science:rivers: rivers: idea: son : myth:deity: deity: 15 : deit: god
Geography: red : pass :ers. :nation : is : lies: lies :lies :ridge: 10
Other:nger :(rl: : (rl:: (rl: : game: you : use : team: film: band
Biology:rine : dna : cell:plant: grow:ase, :some :genes: gene: ten
Fine Arts: solo:piece: opera : st. : opera:opera : chor:pera :opera: oper
Physics: law :zero :less : time : due : can :ying :time :quark: spin
Chemistry: mole: chemis: chemist: sulf:sion :nium : chem: chemi: ion : gas
Mathematics: proof: be :proof: any : has : are : four: set : curv: x
Earth Science: laye: form:rock : rock : or : know: known:land : rock: zone
Science:is\n : law :ine. : atom: 10 : are : this: n : [*] : (*)
Astronomy:lion :cloud: earth: eart: radi: its : mass: mass :mass : star
```

Character *n*-grams

Listing 12 : Character *n*-grams

```
History: khan: case: post: colo: new :orth : died:ses. : i : (+)
Literature: epic:novel: " : play :play : play: poem :poem : poet: poem
Social Science:rivers: rivers: idea: son : myth:deity: deity: 15 : deit: god
Geography: red : pass :ers. :nation : is : lies: lies :lies :ridge: 10
Other:nger :(rl: : (rl:: (rl: : game: you : use : team: film: band
Biology:rine : dna : cell:plant: grow:ase, :some :genes: gene: ten
Fine Arts: solo:piece: opera : st. : opera:opera : chor:pera :opera: oper
Physics: law :zero :less : time : due : can :ying :time :quark: spin
Chemistry: mole: chemis: chemist: sulf:sion :nium : chem: chemi: ion : gas
Mathematics: proof: be :proof: any : has : are : four: set : curv: x
Earth Science: laye: form:rock : rock : or : know: known:land : rock: zone
Science:is\n : law :ine. : atom: 10 : are : this: n : [*] : (*)
Astronomy:lion :cloud: earth: eart: radi: its : mass: mass :mass : star
```

Accuracy: 0.808

What else?

What else?

- Wikipedia categories

What else?

- Wikipedia categories
- How questions are written (quotas per packet)

What else?

- Wikipedia categories
- How questions are written (quotas per packet)
- Syntax (paying attention to words after “this”)

What else?

- Wikipedia categories
- How questions are written (quotas per packet)
- Syntax (paying attention to words after “this”)
- Authors of questions