Start Recording

- Today:
  - Project 3 review
  - Assignment 3 review
  - Evaluation
  - Project 5

Reminders

- Project 4 is due Thursday.
- Writing assignment is due next Tuesday.
- Project 5 is due a week from Thursday.
- Course evaluations available soon (check e-mail)

Project 4

- Please include an output file called output
- I don't care if your condor file stdout goes to output or netid.out
- Please generate your output files in order:
  - Loop through files alphabetically (this should be automatic)
  - Print targets as they are found
  - For extra credit, still print in the order you found them
- How is everybody's run time?
  - Full points if your code runs in about an hour
  - Some points off if it runs in about ten hours
  - More points off if it takes a day or more
  - Tips? Lessons learned on runtime?

General Comments

- Keep your readme somewhat professional
- Use check_homework.sh
- In the 570s don’t hard code input and output

Writing Assignment

- Remember that this paper should conform to achademic writing standards:
  - Use the template provided
  - Plan your paper and use a logical flow: introduction, body, conclusion
  - Write in an academic tone
  - Proof read! Or have a friend proof read it for you.

Self Quiz

- Available for extra practice and self evaluation
- We will go over the solution on the last day of class

Project 3

- A solution is on the website.

```
static void Main(string[] args)
{
    FST fst = new FST();
    // Assumption: args[0] is input, args[1] is output
    using (StreamWriter sw = new StreamWriter(args[1]))
    {
        sw.WriteLine("<meta http-equiv='content-type' content='text html; charset="UTF-8' />");
        foreach (String line in File.ReadAllLines(args[0]))
            sw.WriteLine(fst.Breaker(line) + "<br>");
        sw.WriteLine("");
    }
}
```

```
class FST
{
    HashSet<char> V1 = new HashSet<char>("เแโใไ");
    HashSet<char> C1 = new HashSet<char>("กขคฆงจฉชซฌญฎฏฐฑฒณดตถทธนบปฝพฟภมยรฤลฦวศษสหฬอฮ");
    HashSet<char> C2 = new HashSet<char>("รลวนม");
    HashSet<char> V2 = new HashSet<char>("ี");
    HashSet<char> T = new HashSet<char> { '\u0E48', '\u0E49', '\u0E4A', '\u0E4B' };
    HashSet<char> V3 = new HashSet<char>("าอยว");
    HashSet<char> C3 = new HashSet<char>("งนมดบกยว");
    int state = 0;

    public String Breaker(string input)
    {
        state = 0;
        StringBuilder output = new StringBuilder();
        for (int i = 0; i < input.Length; i++ )
        {
            char c = input[i];
            if (state == 0)
            {
                output.Append(c);
                if (V1.Contains(c))
                    state = 1;
                else if (C1.Contains(c))
                    state = 2;
                else
                    state = -1; //fail state
            }
            else if (state == 1)
            {
                output.Append(c);
                if (C1.Contains(c))
                    state = 2;
                else
                    state = -1;
            }
```

```
else if (state == 2)
{
    output.Append(c);
    if (C2.Contains(c))
        state = 3;
    else if (V2.Contains(c))
        state = 4;
    else if (T.Contains(c))
        state = 5;
    else if (V3.Contains(c))
        state = 6;
    else if (C3.Contains(c))
        state = 9;
    else if (V1.Contains(c))
        state = 7;
    else if (C1.Contains(c))
        state = 8;
    else
        state = -1;
}
else if (state == 3)
{
    output.Append(c);
    if (V2.Contains(c))
        state = 4;
    else if (T.Contains(c))
        state = 5;
    else if (V3.Contains(c))
        state = 6;
    else if (C3.Contains(c))
        state = 9;
    else
        state = -1;
}
else if (state == 4)
{
    output.Append(c);
    if (T.Contains(c))
        state = 5;
    else if (V3.Contains(c))
        state = 6;
    else if (C3.Contains(c))
        state = 9;
    else if (V1.Contains(c))
        state = 7;
    else if (C1.Contains(c))
        state = 8;
    else
        state = -1;
}
```

```
else if (state == 5)
{
    output.Append(c);
    if (V3.Contains(c))
        state = 6;
    else if (C3.Contains(c))
        state = 9;
    else if (V1.Contains(c))
        state = 7;
    else if (C1.Contains(c))
        state = 8;
    else
        state = -1;
}
else if (state == 6)
{
    output.Append(c);
    if (C3.Contains(c))
        state = 9;
    else if (V1.Contains(c))
        state = 7;
    else if (C1.Contains(c))
        state = 8;
    else
        state = -1;
}
else if (state == 7)
{
    state = 1;
    output.Insert(output.Length - 1, " ");
    i--; //don't consume an input
}
else if (state == 8)
{
    state = 2;
    output.Insert(output.Length - 1, " ");
    i--; //don't consume an input
}
else if (state == 9)
{
    state = 0;
    output.Append(" ");
    i--; //don't consume an input;
}
```

Common Error

- Consuming a non-existent input at the end of the line
- Or emitting the output before consuming an input

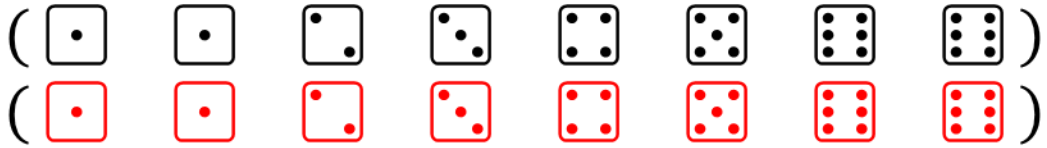
output: ยิน ดี ที่ ได้ รู้ จัก
character: ค
state: 0
output: ยิน ดี ที่ ได้ รู้ จัก ค
character: ุ
state: 2
output: ยิน ดี ที่ ได้ รู้ จัก คุ
character: ณ
state: 4
output: ยิน ดี ที่ ได้ รู้ จัก คุณ

Different Approaches?

- Anyone want to share a different approach
- Lessons learned the hard way?

Assignment 3

Consider weighted dice—one white, and one red. For each die, ⚀ and ⚅ are twice as likely to show as the other four values. What is the probability that the total showing on the two dice will be 7?



The cartesian product has 64 cases.

Ways to get 7:	(1,6)	(2,5)	(3,4)	(4,3)	(5,2)	(6,1)
Number of tuples:	4	1	1	1	1	4

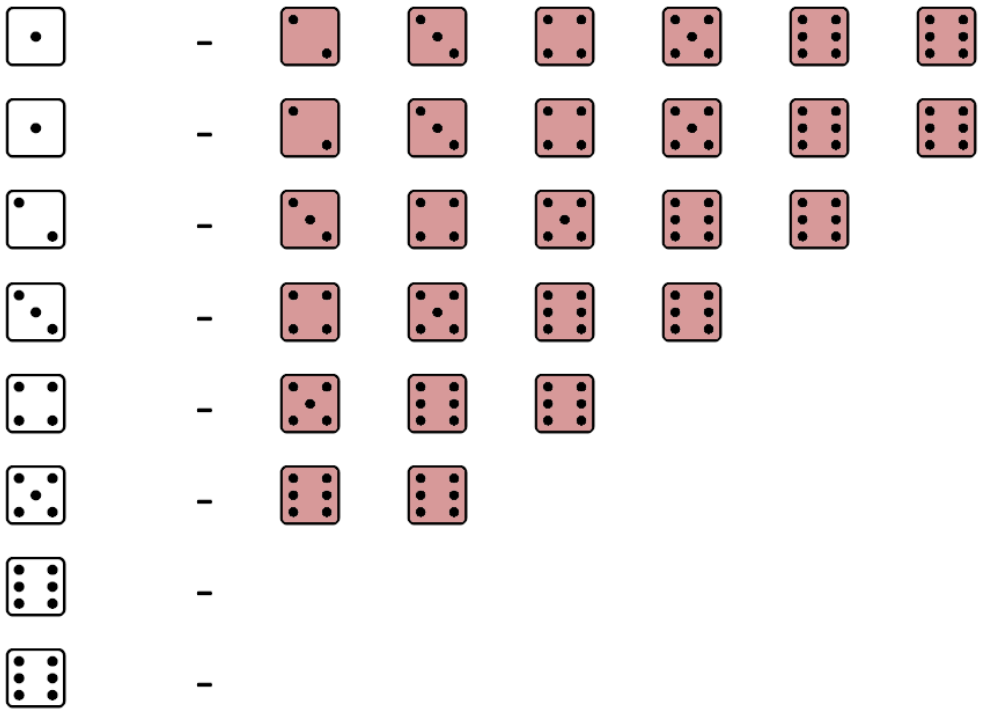
$$\frac{12}{64} = \frac{3}{16} = .1875$$

What is the probability that the total showing on the two dice will be 9 or higher?

(3,6)	(4,5)	(4,6)	(5,4)	(5,5)	(5,6)	(6,3)	(6,4)	(6,5)	(6,6)
2	1	2	1	1	2	2	2	2	4

$$\frac{19}{64} = .296875$$

What is the probability that the red die will show a higher number than the white one?



There are 26 cases

$$\frac{26}{64} = \frac{13}{32} = .40625$$

How many bigrams does the sample contain?

$$158 - 1 = 157$$

$$P(. \mid \text{NN})$$

PRP	VBD	DT	JJ	NN	WP	VBD	RB	IN	DT	NN	IN	DT	NNP	NNP	CC	PRP	VBD	VBN	CD	NNS	RB	IN	VBG	DT	NN	.
he	was	an	old	man	who	fished	alone	in	a	skiff	in	the	gulf	stream	and	he	had	gone	eighty-four	days	now	without	taking	a	fish	.

IN	DT	JJ	CD	NNS	DT	NN	VBD	VBN	IN	PRP	.	CC	IN	CD	NNS	IN	DT	NN	DT	NN	POS	NNS	VBD	VBN	PRP	IN	DT
in	the	first	forty	days	a	boy	had	been	with	him	.	but	after	forty	days	without	a	fish	the	boy	's	parents	had	told	him	that	the

JJ	NN	VBD	RB	RB	CC	RB	VBN	,	WDT	VBZ	DT	JJ	NN	IN	JJ	,	CC	DT	NN	VBD	VBN	IN	PRP$	NNS	IN
old	man	was	now	definitely	and	finally	salao	,	which	is	the	worst	form	of	unluck	,	and	the	boy	had	gone	at	their	orders	in

DT	NN	WDT	VBD	CD	JJ	NN	DT	JJ	NN	.	PRP	VBD	DT	NN	JJ	TO	VB	DT	JJ	NN	VB	IN	DT	NN	IN	PRP$
another	boat	which	caught	three	good	fish	the	first	week	.	it	made	the	boy	sad	to	see	the	old	man	come	in	each	day	with	his

NN	JJ	CC	PRP	RB	VBD	IN	TO	VB	PRP	VB	DT	DT	VBD	NNS	CC	DT	NN	CC	NN	CC	DT	NN	WDT	VBD
skiff	empty	and	he	always	went	down	to	help	him	carry	either	the	coiled	lines	or	the	gaff	and	harpoon	and	the	sail	that	was

VBD	IN	DT	NN	.	DT	NN	VBD	VBN	IN	NN	NNS	CC	,	VBD	,	PRP	VBD	IN	DT	NN	IN	JJ	NN	.
furled	around	the	mast	.	the	sail	was	patched	with	flour	sacks	and	,	furled	,	it	looked	like	the	flag	of	permanent	defeat	.

$$\frac{4}{24} = \frac{1}{6} = .1667$$

$P(\text{DT JJ})$

"How common is the bigram DT JJ in the sample?"

PRP	VBD	DT	JJ	NN	WP	VBD	RB	IN	DT	NN	IN	DT	NNP	NNP	CC	PRP	VBD	VBN	CD	NNS	RB	IN	VBG	DT	NN	.
he	was	an	old	man	who	fished	alone	in	a	skiff	in	the	gulf	stream	and	he	had	gone	eighty-four	days	now	without	taking	a	fish	.

IN	DT	JJ	CD	NNS	DT	NN	VBD	VBN	IN	PRP	.	CC	IN	CD	NNS	IN	DT	NN	DT	NN	POS	NNS	VBD	VBN	PRP	IN	DT
in	the	first	forty	days	a	boy	had	been	with	him	.	but	after	forty	days	without	a	fish	the	boy	's	parents	had	told	him	that	the

JJ	NN	VBD	RB	RB	CC	RB	VBN	,	WDT	VBZ	DT	JJ	NN	IN	JJ	,	CC	DT	NN	VBD	VBN	IN	PRP$	NNS	IN
old	man	was	now	definitely	and	finally	salao	,	which	is	the	worst	form	of	unluck	,	and	the	boy	had	gone	at	their	orders	in

DT	NN	WDT	VBD	CD	JJ	NN	DT	JJ	NN	.	PRP	VBD	DT	NN	JJ	TO	VB	DT	JJ	NN	VB	IN	DT	NN	IN	PRP$
another	boat	which	caught	three	good	fish	the	first	week	.	it	made	the	boy	sad	to	see	the	old	man	come	in	each	day	with	his

NN	JJ	CC	PRP	RB	VBD	IN	TO	VB	PRP	VB	DT	DT	VBD	NNS	CC	DT	NN	CC	NN	CC	DT	NN	WDT	VBD
skiff	empty	and	he	always	went	down	to	help	him	carry	either	the	coiled	lines	or	the	gaff	and	harpoon	and	the	sail	that	was

VBD	IN	DT	NN	.	DT	NN	VBD	VBN	IN	NN	NNS	CC	,	VBD	,	PRP	VBD	IN	DT	NN	IN	JJ	NN	.
furled	around	the	mast	.	the	sail	was	patched	with	flour	sacks	and	,	furled	,	it	looked	like	the	flag	of	permanent	defeat	.

$$\frac{6}{157} = .0382$$

$P(\text{NN} \mid \text{DT JJ})$

"How often does the unigram NN follow the bigram DT JJ?"

"Out of all the DT JJ bigrams, how many of them are followed by NN?"

PRP	VBD	DT	JJ	NN	WP	VBD	RB	IN	DT	NN	IN	DT	NNP	NNP	CC	PRP	VBD	VBN	CD	NNS	RB	IN	VBG	DT	NN	.
he	was	an	old	man	who	fished	alone	in	a	skiff	in	the	gulf	stream	and	he	had	gone	eighty-four	days	now	without	taking	a	fish	.

IN	DT	JJ	CD	NNS	DT	NN	VBD	VBN	IN	PRP	.	CC	IN	CD	NNS	IN	DT	NN	DT	NN	POS	NNS	VBD	VBN	PRP	IN	DT
in	the	first	forty	days	a	boy	had	been	with	him	.	but	after	forty	days	without	a	fish	the	boy	's	parents	had	told	him	that	the

JJ	NN	VBD	RB	RB	CC	RB	VBN	,	WDT	VBZ	DT	JJ	NN	IN	JJ	,	CC	DT	NN	VBD	VBN	IN	PRP$	NNS	IN
old	man	was	now	definitely	and	finally	salao	,	which	is	the	worst	form	of	unluck	,	and	the	boy	had	gone	at	their	orders	in

DT	NN	WDT	VBD	CD	JJ	NN	DT	JJ	NN	.	PRP	VBD	DT	NN	JJ	TO	VB	DT	JJ	NN	VB	IN	DT	NN	IN	PRP$
another	boat	which	caught	three	good	fish	the	first	week	.	it	made	the	boy	sad	to	see	the	old	man	come	in	each	day	with	his

NN	JJ	CC	PRP	RB	VBD	IN	TO	VB	PRP	VB	DT	DT	VBD	NNS	CC	DT	NN	CC	NN	CC	DT	NN	WDT	VBD
skiff	empty	and	he	always	went	down	to	help	him	carry	either	the	coiled	lines	or	the	gaff	and	harpoon	and	the	sail	that	was

VBD	IN	DT	NN	.	DT	NN	VBD	VBN	IN	NN	NNS	CC	,	VBD	,	PRP	VBD	IN	DT	NN	IN	JJ	NN	.
furled	around	the	mast	.	the	sail	was	patched	with	flour	sacks	and	,	furled	,	it	looked	like	the	flag	of	permanent	defeat	.

$$\frac{5}{6} = .833$$

Estimate $P(\text{DT JJ} \mid \text{NN})$

"How often would we expect to see DT JJ following NN in the corpus, based on the prior probabilities of unigram NN and bigram DT JJ, and the measured conditional probability $P(NN|DT\ JJ)$?"

$$P(DT\ JJ|NN) = \frac{P(NN|DT\ JJ)P(DT\ JJ)}{P(NN)}$$

$$= \frac{\frac{5}{6} \times \frac{6}{157}}{\frac{12}{79}} = \frac{395}{1884} = .20966$$

Note: the observed value in the sample is: $\frac{1}{24} = .042$

A = { *gnat*, *beet* } B = { *loon*, *fee* } C = { *peel*, *pool*, *he*, *sand* }

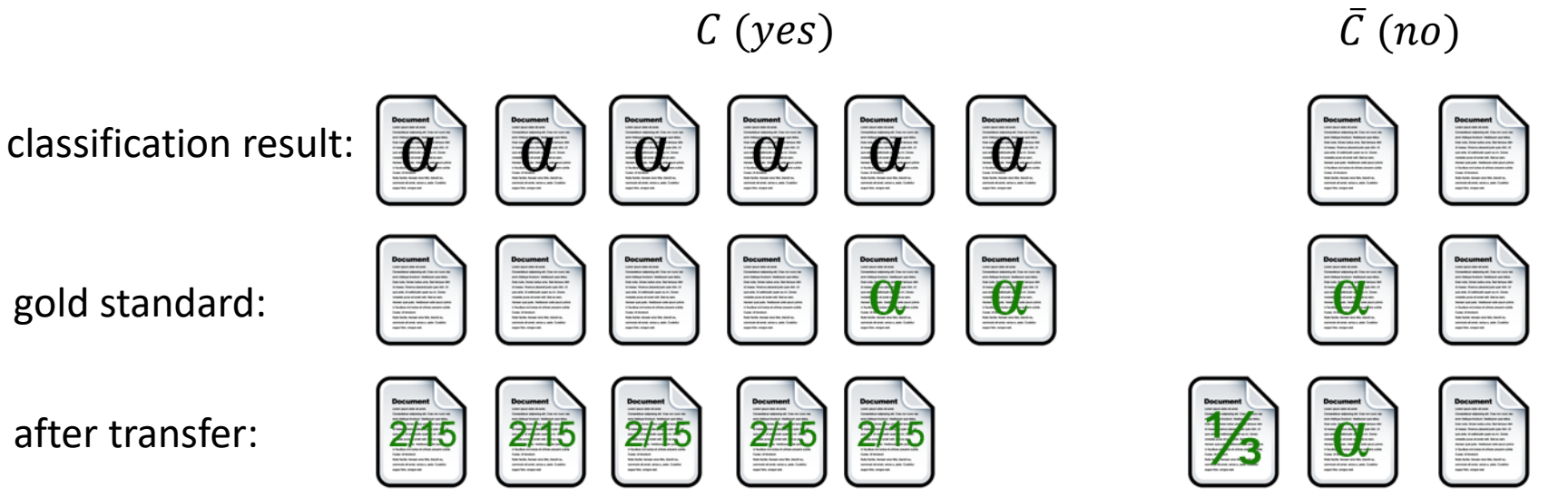
$$P(high|A) = \frac{1}{2}$$

$$P(high|B) = 1$$

$$P(high|C) = \frac{3}{4}$$

$$P(high) = P(high|A)P(A) + P(high|B)P(B) + P(high|C)P(C)$$

$$P(high) = \frac{3}{4}$$



T = { the transferred document actually mentions the promoter }
S = { the final selection actually mentions the promoter }

$P(T) = \frac{1}{3}$

$P(S|T) = \frac{2}{3}$

$P(\bar{T}) = \frac{2}{3}$

$P(S|\bar{T}) = \frac{1}{3}$

$P(S) = P(S|T)P(T) + P(S|\bar{T})P(\bar{T})$

$= \frac{4}{9}$

method 1

$$\left(\frac{1}{3} \times \frac{1}{3}\right) + \left(\frac{1}{3} \times 1\right) + \left(\frac{1}{3} \times 0\right)$$
$$= \frac{4}{9}$$

method 2

$$P(T|S) = \frac{P(S|T)P(T)}{P(S)}$$

$$P(T|S) = \frac{\frac{2}{3} \times \frac{2}{6}}{\frac{4}{9}}$$

$$P(T|S) = \frac{1}{2}$$

Project 5

- Rank language likelihood
  - Assignment is based on word probabilities, but can also be done by characters
  - Existing language models for 15 languages
    - Lists top 1500 most common words for each language
    - Based on these, for each text calculate P(lang|text) for all languages, pick the most probable
  - You will need to use some kind of smoothing algorithm
    - Takes care of unseen words
    - Up to you how to do it, but it should perform well (still label correctly most of the time)
    - Document in writeup how your smoothing works

Project 5

- Rank language likelihood
  - Extra credit (15 points)
    - Add some kind of a threshhold below which you won't pick any language.
    - Run on data that includes unseen languages
    - Report on how well it does in readme

Project 5

- Extra Credit will require an extra project turn-in.
  - Run it via a script run-extra.sh (name the code whatever you like)
  - So the normal files look like this:
    run.sh
    compile.sh
    output
  - The extra credit files look like this:
    run-extra.sh
    compile-extra.sh
    output-extra
  - I've updated the pdf.

Next Time

- Evaluation