COLX 523 Annotation Plan

Bingyang Hou, Jae Ihn, Jiexin Kuang, Jinhong Liu, Min Zeng

---

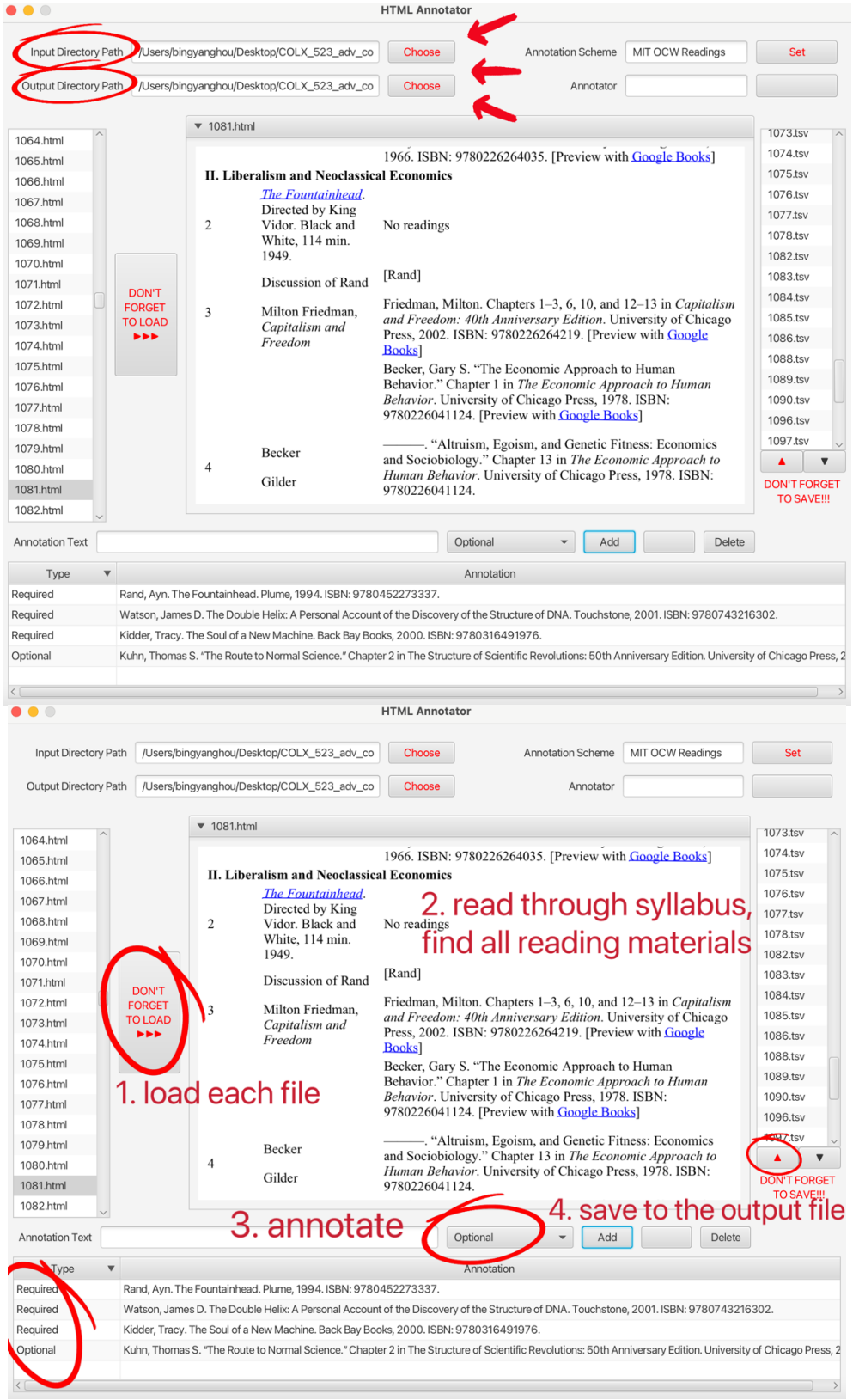
1. **What is the exact kind of annotation you will be doing?**

   - We will be doing binary categorization of reading materials for each course listed in the MIT OpenCourseware website as "required" or "optional". Annotators will be asked to read through each syllabus to:
     - Find reading materials
     - Decide whether the readings are required or optional
   - We defined “reading materials” to include:
     - Books
     - Academic papers
     - Journal articles
     - News articles
   - Our definition of “reading materials” do not include:
     - Websites
     - Videos
     - Podcasts

2. **What tools will you be using to create your annotation?**

   - We will be using a Java program called “HTML Annotator”, created by Jae Ihn for this project.
   - We are not using GATE for this project for a couple of reasons:
     - GATE is a relatively complex tool that requires a significant amount of time and effort to learn and use effectively.
     - GATE can read in HTML files, but completely discards any formatting. Because our corpus documents are made for the web, formatting actually provides a lot of information about where the targeted reading materials might be.
     - GATE is relatively slow—after text is annotation is selected, it takes 1-2 seconds before the annotation window pops up. Since one document can contain many annotations (between 1 to 200), it was crucial for us that the annotation process is fast and painless.
   - The custom HTML Annotator program aims to provide a simple and user-friendly alternative to GATE, and addresses the problems we had above. It was made in a very short time frame (1-2 days), and covers only the basic functions required for this project.
   - A copy of the HTML Annotator can be found in the `milestone_2` repo.
   - We will not be using Mechanical Turk for several reasons:
     - We are using our own annotation tool, which makes distribution (and maintenance of bugs) complicated;
     - Because we will only be annotation part of the corpus (about ¼), we want to maintain high quality for the annotations that we do obtain.

Tutorial: How to use HTML Annotator



- First, the annotators need to choose the input and output directory path. The program loads all .html documents from the input directory path, and exports all annotations to the output directory path as .tsv files.
    - We used the concatenated HTML files for each course as our inputs. These can be found in: `data/concat.zip`
    - We set up a directory inside `data/annotated` for each of the annotators, and used that as our output directories.
- Next, the annotation scheme must be "set". Currently, the annotation scheme and annotator functions are disabled. In the future, we might extend the program so that people can load different annotation schemes and select different annotators. For now, pressing the "set" button will load our annotation labels "required" and "optional" to the drop-down menu below.
- The .html file is selected from the left. The "load" button must be pressed to load its contents into the center pane.
- Text in the central pane can be highlighted using mouse-drag. The highlighted text can be copy-and-pasted, either via right-click or the CTRL+C, CTRL+V keyboard shortcuts.
- After copying the desired segment into the "Annotation Text" text field, we can choose a corresponding label from the drop-down menu. In our case, we choose "required" or "optional". Pressing the "add" button will save the annotation into the table below.
- Rows from the table can be deleted using the "delete" button, in case of incorrect annotations.
- After annotating the current document, the annotation table can be exported as .tsv files using the "up" button (i.e. button in red, with triangle pointing up). Users can see in the right list view the files that are created inside the output directory.
- After the export, the annotation table will be cleared. Users can move onto the next document by loading another file from the left list view.
- Users can use the "up" and "down" button to load existing .tsv files, if they wish to edit their content. This will populate the annotation table with the corresponding files.

3. **Who will be your annotators?**
    - All members of our group will participate in the annotation process.
    - We have assigned batches of documents so that there are at least two people for each annotation.
    - We have also distributed the batches so that we get a variety of course topics, as well as several courses within the same topic.

Member	Batch 1	Batch 2	Batch 3	Batch 4
Ivan	0 - 250	500 - 750	250 - 500	750 - 1000
Jae	1000 - 1250	500 - 750	1250 - 1500	750 - 1000
Biya	1000 - 1250	1500 - 1750	1250 - 1500	1750 - 2000
Jessie	1500 - 1750	2000 - 2250	1750 - 2000	2250 - 2500
CC	0 - 250	2000 - 2250	250 - 500	2250 - 2500
Interannotation total	500	1250	1750	2500

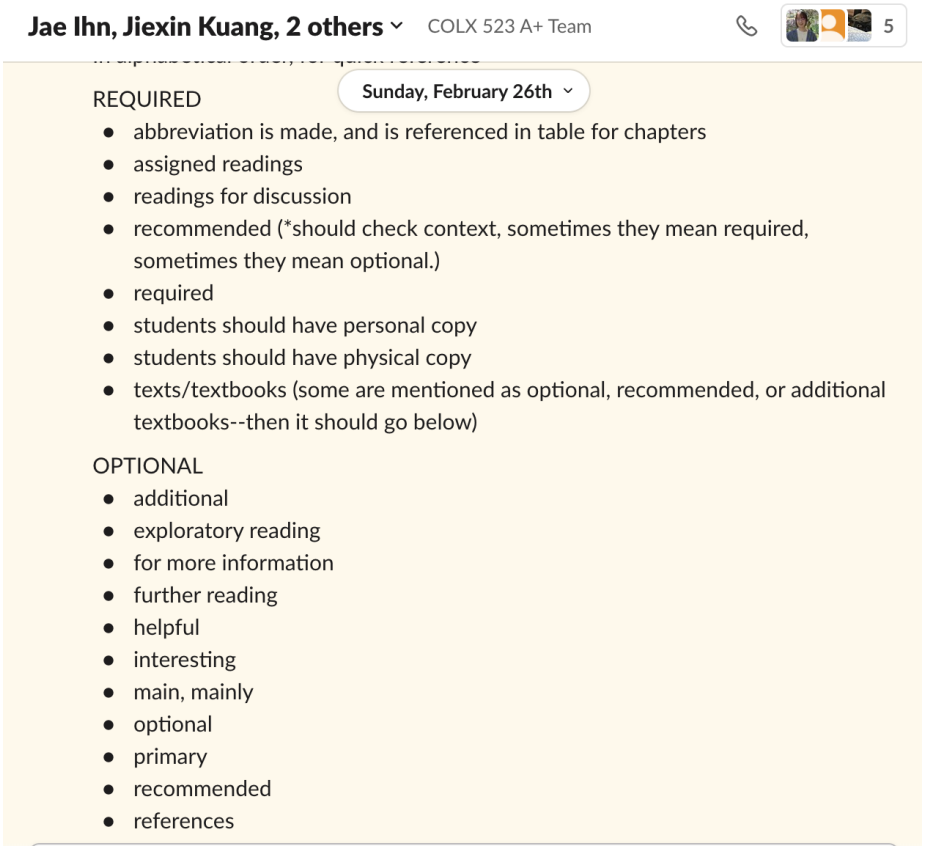
4. **What is the expected amount of data you will be able to have before the next milestone?**

1073

Optional	**Atkinson, A., and J. Stiglitz. Lectures on Public Economics. Mcgraw-Hill College, 1980. ISBN: 9780070841055.**
Optional	Auerbach, A., and M. Feldstein. Handbook of Public Economics. Vol 1. North Holland, 1985. ISBN: 9780444548931.
Optional	Handbook of Public Economics. Vol 2. North Holland, 1987. ISBN: 9780444879080.
Optional	Handbook of Public Economics. Vol 3. North Holland, 2002. ISBN: 9780444823144.
Optional	Handbook of Public Economics. Vol 4. North Holland, 2002. ISBN: 9780444823151.
Optional	Gruber, J. Public Finance and Public Policy. 3rd ed. Worth Publishers, 2009. ISBN: 9781429219495.
Optional	Institute for Fiscal Studies. Dimensions of Tax Design: The Mirrlees Review. Oxford University Press, 2010. ISBN: 9780199553754.
Optional	Myles, G. Public Economics. Cambridge University Press, 1995. ISBN: 9780521497695.
Optional	Slemrod, J., and J. Bakija. Taxing Ourselves: A Citizen's Guide to the Debate over Taxes. MIT Press, 2008. ISBN: 9780262195737.
Optional	U. S. Congress, Congressional Budget Office. The Budget and Economic Outlook: An Update. BiblioGov, 2012. ISBN: 9781249915164.
Required	Chetty, R., A. Looney, et al. "Salience and Taxation: Theory and Evidence." American Economic Review 99, no. 4 (2009): 1145–77.
Required	Congdon, W., J. Kling, and S. Mullainathan. Policy and Choice: Public Finance through the Lens of Behavioral Economics. Brookings Institution Press, 2011. IS
Required	Diamond, P., and D. McFadden. "Some Uses of the Expenditure Function in Public Finance." Journal of Public Economics 5, no. 3–4 (1976): 373–79.
Required	Einav, L., D. Knoepfle, et al. "Sales Taxes and Internet Commerce." NBER Working Paper, no. 18018, April 2012.
Required	Ellison, G., and S. Ellison. "Tax Sensitivity and Home State Preferences in Internet Purchasing." American Economic Journal: Economic Policy 1, no. 2 (2009):
Required	Evans, W., J. Ringel, and D. Stech. "Tobacco Taxes and Public Policy to Discourage Smoking." In Tax Policy and the Economy. Vol 13. Edited by J. Poterba. M
Required	Goulder, L., and R. Williams. "The Substantial Bias from Ignoring General Equilibrium Effects in Estimating Excess Burden, and a Practical Solution." Journal o
Required	Gruber, J., and B. Koszegi. "Tax Incidence When Individuals Are Time-Inconsistent: The Case Of Cigarette Excise Taxes." Journal of Public Economics 88, no.
Required	Harding, M., E. Leibtag, et al. "The Heterogeneous Geographic and Socioeconomic Incidence of Cigarette Taxes: Evidence from Neilsen Homescan Data." (PD
Required	Hausman, J. "Exact Consumers Surplus and Deadweight Loss." American Economic Review 71, no. 4 (1981): 622–76.
Required	Hausman, J., and W. Newey. "Nonparametric Measurement of Exact Consumers Surplus and Deadweight Loss." Econometrica 63 (1995): 1445–76.
Required	Manning, W., Willard, et al. "The Taxes of Sin: Do Smokers and Drinkers Pay Their Way?" Journal of the American Medical Association 261, no. 11 (1989): 160
Required	Marion, J., and E. Muehlegger. "Measuring Illegal Activity and the Effects of Regulatory Innovation: A Study of Diesel Fuel Tax Evasion." Journal of Political Ecc
Required	Salanie, B. Chapter 1 in Economics of Taxation. MIT Press, 2011. ISBN: 9780262016346.
Required	Atkinson, A., and J. Stiglitz. Chapter 6 in Lectures on Public Economics. McGraw Hill College, 1980. ISBN: 9780070841055.
Required	Ballard, C., D. Fullerton, et al. Chapters 2, and 3 in A General Equilibrium Model for Tax Policy Evaluation. University of Chicago Press, 1985. ISBN: 978022603
Required	Bradford, D. "Factor Prices May be Constant but Factor Returns are Not." Economic Letters 1, no. 3 (1978): 199–203.
Required	Cutler, D. "Tax Reform and the Stock Market: An Asset Price Approach." (PDF - 1.6MB) American Economic Review 78, no. 5 (1988): 1107–17.
Required	Fullerton, D., and G. Metcalf. "Tax Incidence." In Handbook of Public Economics. Vol 4. Edited by A. Auerbach, and M. Feldstein. North Holland, 2002. ISBN: 9

- Originally, our naïve ambition was to finish the entire corpus (~2500 documents). Once we started annotating, we quickly realized that we have underestimated how long it takes to annotate one document. Some courses have up to 200 readings!
- We currently expect to have around 700 annotations, each annotated by two different people. This means we will have about 1400 separate annotations.

5. Any steps you have taken to ensure the quality of your annotation?
    - For more information, consult the `annotation_material` file in `milestone_2`.
    - We had a group meeting to train the annotators on how to use the annotation tool and discuss how to categorize the readings correctly.
    - We created an open environment via our group Slack channel, where annotators could ask and discuss any ambiguous cases.
    - We provided regular feedback to each other to help improve the accuracy and consistency in the annotations.
    - We kept a running set of rules in a pinned message in our group chat, such as below:



6. **Did everything go smoothly in the Pilot Study?**
    - We have started our annotation process. Currently we have about 600 single annotations over 300 documents, since we have two annotations per one document.
    - The seemingly simple and easy task—finding reading materials from course syllabi and classifying them as required or optional—turned out to be quite difficult and time consuming.
    - This was not all due to the longer-than-expected annotations, that we discussed above. Each course syllabus is written by the course instructor. Many professors presented their readings in their own unique way. This resulted in many fuzzy cases that made decisions in annotation difficult. We compile some examples in the following pages.

Star (*) or No Star, they are all required…

1087.html

readings

The reading list is intentionally long, to give those of you interested in the field an opportunity to dig deeper into some of the topics in this area. The lectures will cover the material marked with an * in detail and also discuss the material without an *, but in less detail.

First Half of the Class Taught by Prof. Daron Acemoglu

LEC #	TOPICS	READINGS
Voters (Sessions 1–7)		
		Arrow, Kenneth J. (1951, 2nd ed., 1963). *Social Choice and Individual Values*, Yale University Press.
		Black, Duncan (1948). "On the Rationale of Group Decision-making", *Journal of Political Economy* 56 (1), pp. 23-34.
		Downs, Anthony (1957). "An Economic Theory of Political Action in a Democracy", *Journal of Political Economy* 65 (2), pp. 135-150.
		Austen-Smith, David and Jeffrey S. Banks (2000). *Positive Political Theory I: Collective Preference*, University of Michigan

Nothing's "required" per se, but…

1068.html

While there is no required textbook, the course does make frequent use of the following books:

Buy at MIT Press BF = Blanchard, Olivier J., and Stanley Fisher. *Lectures on Macroeconomics*. Cambridge, MA: MIT Press, 1989, ISBN: 0262022834.

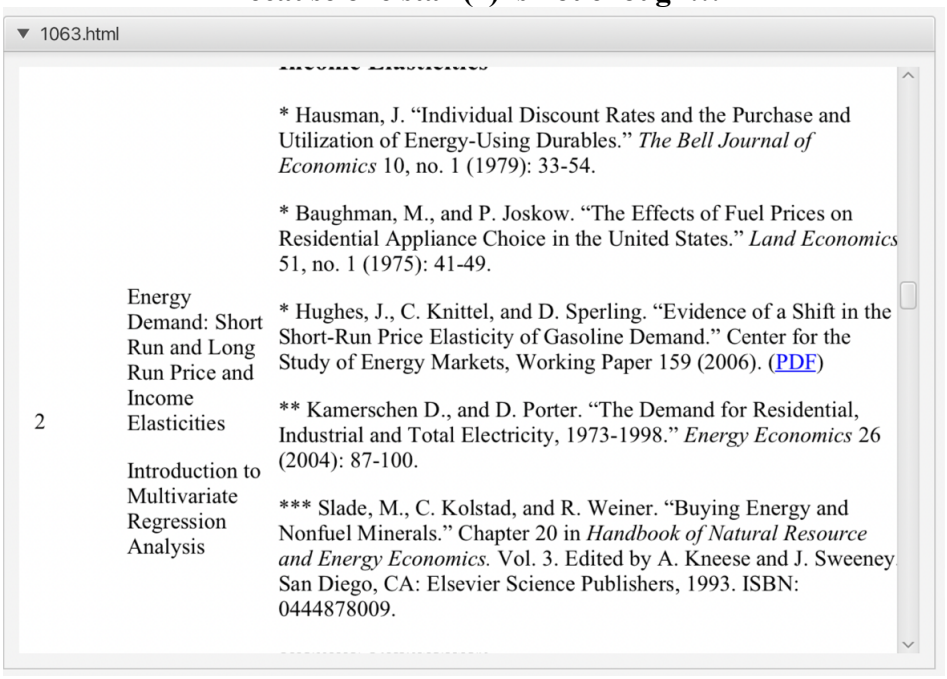
Deaton = Deaton, Angus. *Understanding Consumption*. New York, NY: Oxford University Press, 1992, ISBN: 0198288247.

Buy at MIT Press LS = Ljungqvist, Lars, and Thomas J. Sargent. *Recursive Macroeconomic Theory*. 2nd ed. Cambridge, MA: MIT Press, 2004, ISBN: 026212274X.

Readings by Session

LEC #	TOPICS	READINGS
		Backus, David, Bryan Routledge, and Stanley Zin. "Exotic Preferences for Macroeconomists." NBER Working Paper No. 10597, June 2004.
		Essential
		Deaton. Chapter 1.
		Lucas, Robert E., Jr. *Models of Business Cycles*. Reprint ed. Malden

Because one star (*) is not enough…



Almost, but not quite…

Almost Required Readings

McCloud, Scott. *Understanding Comics: The Invisible Art.* New York, NY: Harper Paperbacks, 1994. ISBN: 9780060976255.

Read the readings to win your own copy…