INSTRUCTION FOR CHILE
By Constanza Schibber

Please read the instructions carefully. I will revise the instructions when you are finished to make sure all the items were addressed. If you have any questions or run into any problems, do not hesitate to email me immediately.

1. SCHEDULE

Instructions Submission: May 10, 2014

Deadline: May 17, 2014

Please provide e-mail updates before the deadline

2. Code Submission - Checklist

a. Include COMMENTS in your code

b. Include a "Read Me" file with instruction on how to run the code

3. Time Expectations

The previous code took 40 hours. As we talked about on Wednesday, I expect you to finish the work in 10 to 15 hours. You should use as much of the previous code as possible.

If you want to start part of the code from scratch it is your responsibility. We have a pre-assigned number of hours for each task and we cannot add another 40 hours to the task.

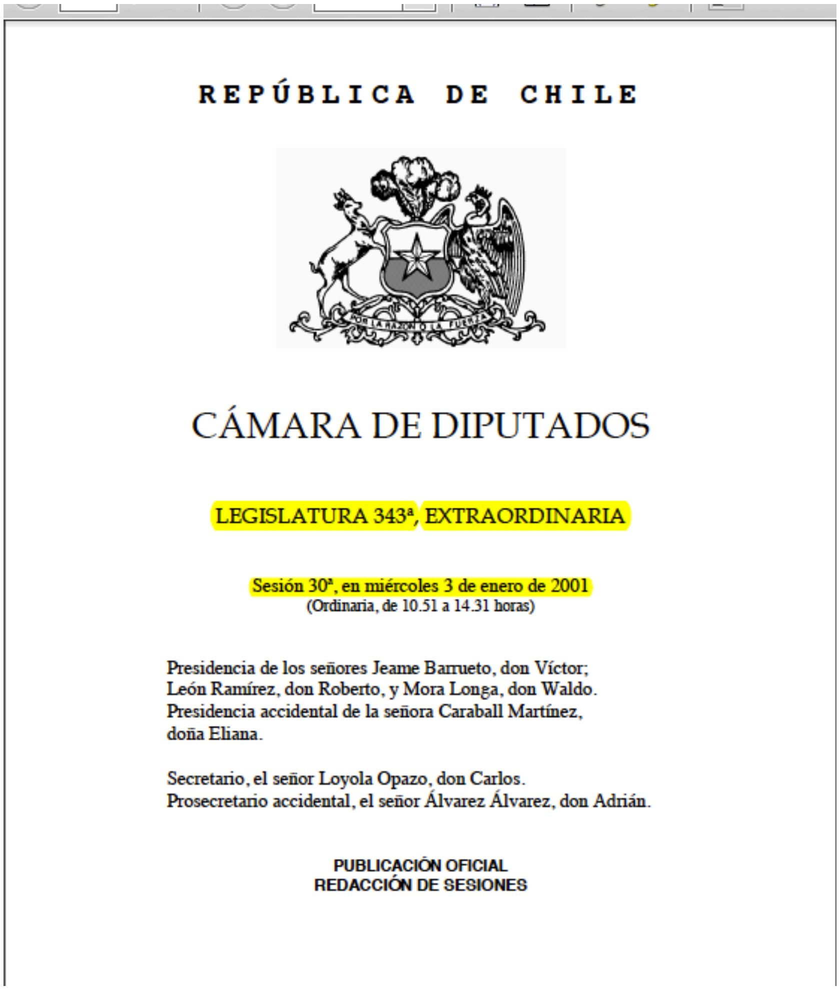
4. Instructions

The instructions are divided in 2 parts. Start by (1) turning the PDF into a text (2) using my function clean_text to take all the *á é* etc. as we talked in the meeting (3) You can collapse the text and erase all the spaces using .strip. You'll have to erase spaces in the regular expressions provided too. BUT you might not need to do it because I made work simpler.

4.1. Part 1 --

This part seems to be already in the code you wrote

From each file we need the following information that appears on PAGE 1. See the highlighted sections in the example below.



REPÚBLICA DE CHILE

CÁMARA DE DIPUTADOS

LEGISLATURA 343ª, EXTRAORDINARIA

Sesión 30ª, en miércoles 3 de enero de 2001
(Ordinaria, de 10.51 a 14.31 horas)

Presidencia de los señores Jeame Barrueto, don Víctor;
León Ramírez, don Roberto, y Mora Longa, don Waldo.
Presidencia accidental de la señora Caraball Martínez,
doña Eliana.

Secretario, el señor Loyola Opazo, don Carlos.
Prosecretario accidental, el señor Álvarez Álvarez, don Adrián.

PUBLICACIÓN OFICIAL
REDACCIÓN DE SESIONES

(1) From the example we need to keep a record of "LEGISLATURA 343" -- So, look from "LEGISLATURA" and get everything until the comma. In Betul's intruction this was called **LEGISLATURE_NUMBER**

(2) Get the word after the comma. Here the word is "EXTRAORDINARIA". In Betul's instructions this is **LEGISLATURE_TYPE**

(3) Phrase "Sesión 30" needs to be saved -- Look for "Sesión" and get everything until the comma. In Betul's instructions this is called **SESSION_NUMBER**

(4) Phrase “en miércoles 3 de enero de 2001“ has to be saved-- this one is the date organized as: *miércoles* is the day of the week, *3* is the day, *enero* is the month, *2001* is the year. In Betul’s intruction this is **DATE.**

I think there is more detailed code in what you did to transform the date into something more readable, like mm/dd/yyyy. For now extract the phrase. We’ll focus on this at the end.

Put all this information in a list. Let’s call it **SESSION_INFO** for the purpose of the instructions. You can call it whatever you want.

4.2 Part II

The following is a bit more SIMPLE than the instructions given by Betul

I provide a list of a sample of legislators in Chile. I will provide the full list when the code is ready.

Use the variable NAME which contains the names of the legislators. You will use their names to record how they voted. Legislators can vote “YES”, “NO”, “ABSTAIN”, “ABSENT”.

The final product should be a CSV file, looking like this:

Name	Vote 1	Vote 2
Name 1	YES	ABSTAIN
Name 2	NO	ABSENT
Name 3	YES	YES
...	...	...

--- Step 1---- Dividing the text

The phrases that I will provide to find the VOTES need to be search within sections of the text. To divide the text make a list with information provided in the Contents Table of the file. The content table is in the first pages of the file. It is easy to find. For example:

V.Orden del Día.

- Normativa contra la evasión tributaria. Primer trámite constitucional. (Continuación) 8

- Creación de la Defensoría Penal Pública. Tercer trámite constitucional......... 64
- Concesión de nacionalidad por gracia al sacerdote Pierre Dubois 64

VI.**Proyectos de acuerdo.**

- Condonación de deudas de la ex Corporación de Reforma Agraria, Cora. (Votación)... 66
- Fijación de salvaguardia para el queso 66
- Suspensión de envío a Dicom de datos de deudores morosos de Inverca e investigación de relaciones entre esta empresa y el Ministerio de Vivienda 69

VII.Incidentes.

- Demora de la Corporación Nacional de Desarrollo Indígena en resolver asunto de su competencia

....

Create a LIST IN LIST:

Basically, create a single list, but the list contains lists.

Look for “Orden del dia”. Each item in the table of contents has to be stored. Then look for “Proyectos de acuerdo”. Items here end with word “Incidentes”

For instance, from the example I pasted above, the list should look like this:

[“Orden del dia”, “Normativa contra la evasión tributaria. Primer trámite constitucional. (Continuación)”], [“Orden del dia”, Creación de la Defensoría Penal Pública. Tercer trámite constitucional], [“Proyectos de acuerdo”, Condonación de deudas de la ex Corporación de Reforma Agraria, Cora. (Votación)]

Then you will loop within the list, specially for the 2nd item of the sublist. Search for the phrase (you can make the phrase shorter by looking for the first sentence in each phrase; it might work better). We need to do STEP 2 after each phrase BUT before the phrase in the following sublist starts. This is a way of dividing the text and being more efficient about doing STEP 2.

--- Step 2 --- Regular Expressions

Look for the phrase “Votaron por la afirmativa los siguientes” OR “Voto por la afirmativa”. If individuals after the phrase (but before the phrases below) are matched to a name in the list I gave you, they are assigned a YES.

Then look for the phrase “Votaron por la negativa los siguientes” OR “Voto por la negativa”. The individuals after the phrase (but before the phrases below) are assigned a NO.

Then look for the phrase “Se abstuvieron” OR “Se abstuvo”. The individuals after the phrase are assigned ABSTAIN.

Else, the individuals are assigned “ABSENT”.

-- Basically, you need to create an empty list. Loop over the names I provided. If the name matches the name in the text, append “YES” to the empty list, if the name doesn’t match, continue. If the name matches the 2nd phrase, append “NO”, else continue. If the name matches the 3rd phrase, append “ABSTAIN”, else continue and finally append “ABSENT”.

If NONE OF the phrases appear in the section of the text, append “NO VOTE” to the sublist you are using (the one w/the phrases).
If ANY OF the phrases appear in the section of the text, append “VOTE” to the sublist you are using.

--- STEP 3 --- SAVING --

There are 2 csv files:

---- File 1

Save Step 2 as a row in a csv file
Basically after each loop in Step 2 you save in file 1.

---- File 2

Append SESSION_INFO to each sublist from STEP 1.

Save each sublist as a row in a csv file.

When you finish with a PDF file you save in this save. So before opening a new file you need to save.

ADDITIONAL PROBLEM

Search for the phrases in the text file you create by hand to check out the format. I found that the text file I had, had the list of names BEFORE the phrase. This was some problem in the transformation of the PDF to a TXT. See:

senores diputados: ==> POSSIBLE PHRASE TO SEARCH NEXT IN THE PREVIOUS PARAGRAPH

Acuna, AlvarezSalamanca, Rozas (dona Maria), Caraball (dona Eliana), Ceroni, Coloma, Cornejo (don Patricio), Correa, Delmastro, Encina, Espina, Fossa, Garcia (don Rene Manuel), Garcia (don Jose), GarciaHuidobro, Gonzalez (dona Rosa), Gutierrez, Jaramillo, Jarpa, Jimenez, Leay, Letelier (don Felipe), Luksic, Melero, Monge, Mora, Mulet, Olivares, Ortiz, Perez (don Jose), Perez (don Anibal), Perez (dona Lily), Salas, Sanchez, Valenzuela, Van Rysselberghe, Vargas, Velasco, Venegas y Villouta. ===> NAMES

Votaron por la afirmativa los siguientes [PHRASE PROVIDED]